

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Advancing Automated AMD Stage Grading with 2D and 3D Retinal OCTA Imaging

Permalink

<https://escholarship.org/uc/item/8zc6x3vz>

Author

Zhang, Haochen

Publication Date

2025

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Advancing Automated AMD Stage Grading with 2D and 3D Retinal OCTA Imaging

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Haochen Zhang

Committee in charge:

Professor Truong Nguyen, Chair
Professor Cheolhong An, Co-Chair
Professor Pamela Cosman
Professor William Freeman

2025

Copyright
Haochen Zhang, 2025
All rights reserved.

The dissertation of Haochen Zhang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2025

TABLE OF CONTENTS

Dissertation Approval Page	iii
Table of Contents	iv
List of Figures	vii
List of Tables	ix
Acknowledgements	x
Vita	xii
Abstract of the Dissertation	xiv
Chapter 1 Introduction	1
1.1 Motivation and Objective	1
1.2 Challenges and Contributions	3
Chapter 2 Preliminaries and Dataset	7
2.1 Age-related Macular Degeneration	7
2.2 Retinal Imaging Modalities	9
2.2.1 Optical coherence tomography	9
2.2.2 OCT angiography	11
2.3 OCTA Dataset	14
2.3.1 Existing OCTA datasets	14
2.3.2 Our OCTA dataset	16
Chapter 3 Robust AMD Stage Grading with 3D OCTA Volume	23
3.1 Introduction	23
3.2 Related Works	25
3.2.1 OCTA analysis in computer vision	25
3.2.2 AMD diagnosis with deep learning	26
3.2.3 OCTA datasets	27
3.3 Proposed Methods	28
3.3.1 2D classifier based on OCTA projection images	28
3.3.2 Presence of layer segmentation errors	29
3.3.3 Avoid segmentation errors with 3D input	32
3.4 Experimental Results	33
3.4.1 Influence of layer segmentation errors	35
3.4.2 Performance of deep classifier	37
3.4.3 Ablation study	39
3.4.4 Detailed comparison with human expert	40

	3.5 Discussion	44
Chapter 4	Cross-instrument OCTA Translation with Class Condition	45
	4.1 Introduction	45
	4.2 Related Work	47
	4.2.1 OCTA analysis	47
	4.2.2 Neural style transfer technique	48
	4.3 Proposed Methods	49
	4.3.1 Overview of CycleGAN framework	49
	4.3.2 Explicit class condition	51
	4.3.3 Classifier training with CycleGAN	53
	4.4 Experimental Results	54
	4.4.1 Implement details	54
	4.4.2 Classification performance	55
	4.4.3 Visual inspections	58
	4.5 Additional Experiments	58
	4.5.1 Problem of difference in field of view	58
	4.5.2 Backbone searching for Optovue data	60
	4.6 Discussion	61
Chapter 5	Diffusion Model based Retinal Image Synthesis with Class-Condition . .	63
	5.1 Introduction	63
	5.2 Related Work	65
	5.2.1 Deep learning based DR grading	65
	5.2.2 Diffusion model and data synthesis	66
	5.3 Proposed Method	67
	5.3.1 Text-to-image diffusion model finetuning on fundus images	67
	5.3.2 Semantic quality based evaluation and filtering	68
	5.3.3 Finetuning diffusion model with explicit class condition	70
	5.4 Experiments for Data Imbalance Issue	71
	5.4.1 Implement details	71
	5.4.2 Experimental results	72
	5.4.3 Ablation study	74
	5.5 Additional Experiment for Dataset Extension	78
	5.6 Extended Results on OCTA Images for AMD	83
	5.6.1 Channel problem of OCTA images	83
	5.6.2 Diffusion model training	86
	5.7 Discussion	88
Chapter 6	Conclusion and Future Work	91
	6.1 Conclusion	91
	6.2 Future Work	93
	6.2.1 Synthetic image generation improvement	94

6.2.2	Fine-grained classification with transformers	95
6.2.3	Biomarker discovery with model visualization	96
Bibliography		98

LIST OF FIGURES

Figure 1.1:	Comparison between fundus and OCTA w.r.t. AMD stages.	2
Figure 2.1:	Vision with different types of AMD, from All About Vision website [1]. AMD doesn't cause complete blindness, but losing your central vision can make it harder to see faces, read, drive, or do close-up work like cooking or fixing things around the house.	8
Figure 2.2:	3D view of OCT and OCTA raw volume, modified from Shutterstock [2]. The 3D OCT and OCTA data consist of many B-scans shown as blue box. If we use x,y as spatial direction and z for depth, each B-scan has x and z two dimensions. They are concatenated along y axis to form a 3D volume. . . .	10
Figure 2.3:	Various retinal imaging techniques for diagnosing AMD.	13
Figure 2.4:	Examples of OCTA images in active stage (group1) and normal stage (group4) of AMD captured using Heidelberg and Optovue instruments respectively. The images in active stage illustrate CNV in the avascular slab for Heidelberg and in the outer retina and choriocapillaris slab for Optovue.	18
Figure 3.1:	The proposed network structures for (a) 2D projections, (b) 3D volumes and (c) 3D volumes with 2D projection supervision. The layers in blue have pretrained weights while those in red are trained from scratch.	28
Figure 3.2:	Illustration of the interrelationships among OCT and OCTA raw volume, B-scans, and OCTA projection.	29
Figure 3.3:	Examples of avascular projection w/o and w/ manual layer segmentation error correction by human experts. Layer segmentation errors lead to incorrect vascular networks, missing vessels and noise in OCTA projections, which complicates the classification for both ophthalmologists and neural networks.	31
Figure 3.4:	Illustration of the datasets used in our experiments	34
Figure 3.5:	Standardized reference images used in human expert evaluation experiment.	41
Figure 3.6:	Confusion matrix comparison between our proposed method and human experts on different test sets.	42
Figure 4.1:	Illustration of our framework utilizing CycleGAN to transfer samples between Heidelberg and Optovue domains. For simplicity, only one channel of OCTA projection is displayed, and the identity loss is omitted. For more details please refer to Figure 4.2	50
Figure 4.2:	The disassembly diagram of Figure 4.1 showing the blue (a) and green (b) cycles, respectively, with the proposed supervised (a) and unsupervised (b) class constraints.	51
Figure 4.3:	Illustration of the CycleGAN-based data augmentation for DNN classifier training. Synthetic samples are generated via CycleGAN as shown in Figure 4.2. The classifier can be trained with real or synthetic data from each domain, employing predetermined sampling strategies.	53

Figure 4.4:	Visual inspection of CycleGAN results. Each row shows one example. As can be observed, CycleGAN successfully converts the image style while preserving the vascular patterns.	57
Figure 5.1:	Overview of our text-to-image diffusion model finetuning framework. . . .	68
Figure 5.2:	Comparison of confusion matrices without and with high-semantic-quality synthetic samples. Oversampling helps address data imbalance, while our proposed data synthesis method further enhances performance.	73
Figure 5.3:	Visual inspection of synthetic samples.	74
Figure 5.4:	Classification performance variation w.r.t the amount of synthetic data. . .	75
Figure 5.5:	Classification performance variation w.r.t the choices of semantic score threshold.	77
Figure 5.6:	Confusion matrices corresponding to different oversampling weight configurations. We use $\{\frac{1}{n_i}\}_{i=1}^5$ as weight, so a smaller number in figure representing a larger sampling weight for this class.	78
Figure 5.7:	Illustration of uniform and adaptive synthetic sampling strategies.	79
Figure 5.8:	Classification performance with respect to varying amounts of synthetic data generated by uniform and adaptive sampling strategies.	80
Figure 5.9:	Comparison of Accuracy and Balanced Accuracy variations with increasing amounts of synthetic data under different sampling strategies.	81
Figure 5.10:	Confusion matrices corresponding to different oversampling weight configurations. We use $\{\frac{1}{n_i}\}_{i=1}^5$ as weight, so a smaller number in figure representing a larger sampling weight for this class.	82
Figure 5.11:	Comparison of real and generated avascular OCTA projection.	84
Figure 5.12:	Performance w.r.t semantic quality threshold in content of Heidelberg OCTA data synthesis.	87
Figure 5.13:	Performance w.r.t OCTA synthetic data generated by different versions of proposed diffusion model in content of Heidelberg OCTA data synthesis. .	88
Figure 6.1:	Examples of AblationCAM visualization results.	96

LIST OF TABLES

Table 2.1:	Comparison of public OCTA datasets with the dataset collected by us. Please note that our dataset is specifically curated for AMD stage grading and has the largest number of total and AMD samples available.	15
Table 2.2:	Data statistics of our Heidelberg and Optovue datasets with clinical explanation.	20
Table 3.1:	Distribution of error-free and error-prone samples and associated error rates.	30
Table 3.2:	Classification accuracy with different training/testing datasets. ‘Clean’ means a set with no segmentation errors and ‘Mixed’ means a set mixed with samples with and without errors.	36
Table 3.3:	Ensemble accuracy (%) and RoC-AUC performance of different AMD graders with 2D inputs. Error-free and Error-prone are two testing sets w/ and w/o segmentation errors, respectively. MM: Multimodal information (including OCT B-scan, OCT and OCTA projections), PT: Pretraining.	38
Table 3.4:	Ensemble accuracy (%) and RoC-AUC performance of different AMD graders with 3D inputs. PT: Pretraining, PS: Projection Supervision.	38
Table 3.5:	Ablation experiments w.r.t the choice of channel axis and the loss weight ratio. The accuracy here pertains to the performance of individual classifier trained on the first validation fold, instead of the outcome of ensemble.	40
Table 4.1:	Classification Accuracy without CycleGAN augmentation. Real_ $\{H,O\}$ means a mixture of Real_H and Real_O. Real_H 3D, which uses 3D data as input, is provided for reference.	56
Table 4.2:	Classification Accuracy with CycleGAN augmentation. Please refer to Figure 4.3 for example of different training sample type.	56
Table 4.3:	Classification Accuracy without FoV alignment, which means Optovue samples has doubled FoV than Heidelberg samples but share same resolution.	59
Table 4.4:	Neural network backbone searching result. EfficientNet-b5 is still the optimal choice considering training speed and computational resource constraints.	60
Table 5.1:	Classification performance of LANet with different backbones using the proposed data synthesis on the DDR dataset.	72
Table 5.2:	Performance comparison with real and synthetic data.	76
Table 5.3:	Classification accuracy with different OCTA projections.	85
Table 5.4:	Per-class classification accuracy on real and synthetic validation set. The model is trained on a mixture of real and synthetic data.	89

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Professor Truong Nguyen, Professor Cheolhong An and Professor William Freeman for their invaluable guidance and unwavering support as my committee chair and co-chair. Their expertise and mentorship have played a crucial role in shaping my PhD journey. I also sincerely thank my committee members for their insightful feedback and encouragement throughout my dissertation work. I would like to acknowledge the members of the Video Processing Lab and the Jacobs Retina Center for their technical support and collaborative spirit. Special thanks to all my co-authors and collaborators, whose dedication made our research possible. Finally, I am profoundly thankful to my family and friends for their love, patience, and encouragement. Their support has been a constant source of strength.

Chapter 2 and Chapter 3, in part, are a reprint of the material as it appears in 2023 International Conference on Computer Vision (ICCV) Workshop [3], Haochen Zhang, Anna Heinke, Carlo Miguel B. Galang, Daniel N. Deussen, Bo Wen, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Cheolhong An. “Robust AMD Stage Grading with Exclusively OCTA Modality Leveraging 3D Volume”, in *ICCV* workshop, 2023. The dissertation author is the primary author of this paper. They, in part, are also a reprint of the material as it appears in *RETINA* [4], Anna Heinke, Haochen Zhang, Daniel N. Deussen, Carlo Miguel B. Galang, Alexandra Warter, Fritz G. P. Kalaw, Dirk-Uwe G. Bartsch, Lingyun Cheng, Cheolhong An, Truong Q. Nguyen, and William R. Freeman. “Artificial Intelligence for OCTA-based Disease Activity Prediction in Age-Related Macular Degeneration”, *RETINA*, 2022. The dissertation author is a co-primary contributor to this work responsible for AI classifier development and experiment results analysis.

Chapter 4, in part, is a reprint of the material as it appears in the Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBC) [5], Haochen Zhang, Anna Heinke, Krzysztof Broniarek, Carlo Miguel B. Galang, Daniel N. Deussen, Ines D. Nagel, Katarzyna Michalska-Malecka, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Che-

olhong An. “OCTA-based AMD Stage Grading Enhancement via Class-Conditioned Style Transfer”, in *EMBC*, 2024. The dissertation author is the primary author of this paper. It, in part, is also a reprint of the material as it appears in Scientific Reports [6], Anna Heinke, Haochen Zhang, Krzysztof Broniarek, Katarzyna Michalska-Małecka, Wyatt Elsner, Carlo Miguel B. Galang, Daniel N. Deussen, Alexandra Warter, Fritz G. P. Kalaw, Ines D. Nagel, Akshay Agnihotri, Nehal N. Mehta, Julian Elias Klaas, Valerie Schmelter, Igor Kozak, Sally L. Baxter, Dirk-Uwe G. Bartsch, Lingyun Cheng, Cheolhong An, Truong Q. Nguyen, and William R. Freeman. “Cross-instrument optical coherence tomography-angiography (OCTA)-based prediction of age-related macular degeneration (AMD) disease activity using artificial intelligence”, *Scientific Reports*, 2024. The dissertation author is a co-primary contributor to this work responsible for AI classifier development and experiment results analysis.

Chapter 5, in part, has been submitted for publication of the material as it may appear in International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Haochen Zhang, Anna Heinke, Ines D. Nagel, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Cheolhong An, “Class-Conditioned Image Synthesis with Diffusion for Imbalanced Diabetic Retinopathy Grading” in *MICCAI*, 2025. The dissertation author is the primary author of this submission.

VITA

2013 - 2017	Bachelor in Electronic Engineering, University of Science and Technology of China
2017 - 2020	Master in Electronic Engineering, University of Science and Technology of China
2020 - 2025	Ph. D. in Electrical Engineering, University of California San Diego

PUBLICATIONS

Heinke, A., **Zhang, H.**, Broniarek, K., Michalska-Małeck, K., Elsner W., Galang, C.M.B., Deussen, D.N., Warter, A., Kalaw, F.G.P., Nagel, I.D., Agnihotri, A., Mehta, N. N., Klaas J. E., Schmelter V., Kozak I., Baxter S. L., Bartsch, D.U.G., Cheng, L., An, C., Nguyen, T.Q., and Freeman, W.R., “Cross-instrument optical coherence tomography-angiography (OCTA)-based prediction of age-related macular degeneration (AMD) disease activity using artificial intelligence.”, *Scientific Reports*, 14(1), 27085, 2024.

Zhang, H., Heinke, A., Broniarek, K., Galang, C.M.B., Deussen, D.N., Nagel, I.D., Michalska-Małeck, K., Bartsch, D.U.G., Freeman, W.R., Nguyen, T.Q. and An, C., “OCTA-based AMD Stage Grading Enhancement via Class-Conditioned Style Transfer.”, In *Proceedings of International Conference of the IEEE Engineering in Medicine and Biology Society*, pp.1-5, 2024.

Zhang, H., Heinke, A., Galang, C.M.B., Deussen, D.N., Wen, B., Bartsch, D.U.G., Freeman, W.R., Nguyen, T.Q. and An, C., “Robust AMD Stage Grading with Exclusively OCTA Modality Leveraging 3D Volume.”, In *ICCV workshop*, pp. 2411-2420, 2023.

Heinke, A., **Zhang, H.**, Deussen, D.N., Galang, C.M.B., Warter, A., Kalaw, F.G.P., Bartsch, D.U.G., Cheng, L., An, C., Nguyen, T. and Freeman, W.R., “Artificial Intelligence for OCTA-based Disease Activity Prediction in Age-Related Macular Degeneration”, *RETINA*, pp.465-474, 2022.

Papers under review

Zhang, H., Heinke, A., Nagel, I.D., Bartsch, D.U.G., Freeman, W.R., Nguyen, T.Q. and An, C., “Class-Conditioned Image Synthesis with Diffusion for Imbalanced Diabetic Retinopathy Grading.” (Submitted to *MICCAI 2025*)

Zhang, H., Sinha A., Xu, F., Ma H., Li K., Fan Z., Dai X., Hou T., Zhang P., Vajda P., He Z., “Advancing Multimodal Large Language Models Towards Reasoning based Image Generation.” (Submitted to *ICCV 2025*)

Nagel, I.D., Tran, M.D., **Zhang, H.**, Cheng, L., Heinke, A., Mehta, N.N., Bartsch, D.U.G., Kalaw, F.G.P., Morsy, M., Mueller, A.J., Freeman, W.R., “Reduction of Pigment Epithelium Detachment in Highly Resistant CNV Eyes treated with 6 and 12 mg Faricimab: An OCT-Angiographic and AI-Based Analysis.” (Submitted to *Scientific Reports*)

Papers in preparation

Heinke, A., **Zhang, H.**, Broniarek, K., Michalska-Maleck, K., Arya, M., Nagel, I.D., Mehta, N. N., Morsy, M., Bartsch, D.U.G., Cheng, L., An, C., Nguyen, T.Q., and Freeman, W.R., “AI-Driven Cross-OCTA Device Fine-Grained Classification of AMD Disease Activity using en face OCTA Imaging.” (Experimental results ready, targeting *Scientific Reports*)

Heinke, A., **Zhang, H.**, Broniarek, K., Michalska-Maleck, K., Nagel, I.D., Mehta, N. N., Morsy, M., Bartsch, D.U.G., Cheng, L., An, C., Nguyen, T.Q., and Freeman, W.R., “Impact of Field of View and Image Resolution on AI-Based Detection of AMD Activity in en face OCTA Imaging.” (Experimental results ready, targeting *Translational Vision Science & Technology*)

ABSTRACT OF THE DISSERTATION

Advancing Automated AMD Stage Grading with 2D and 3D Retinal OCTA Imaging

by

Haochen Zhang

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California San Diego, 2025

Professor Truong Nguyen, Chair
Professor Cheolhong An, Co-Chair

Age-related Macular Degeneration (AMD) is a degenerative eye disease that leads to central vision loss, making early detection of AMD pathology, especially choroidal neovascularization (CNV), crucial for effective treatment. Optical Coherence Tomography Angiography (OCTA) emerges as a promising tool for AMD diagnosis by visualizing pathogenic vessels in the subretinal space. However, OCTA's clinical utility is hindered by its early developmental stage. In this thesis, we explore OCTA's potential in AMD diagnosis from perspective view of deep learning-based classifiers, exploiting its capacity to discern hidden patterns and process 3D data efficiently. To our knowledge, this is the first study exclusively utilizing OCTA for AMD stage

grading.

Our research uncovers significant challenges in developing deep learning-based AMD stage graders, including the scarcity of suitable OCTA datasets and variability in image quality due to layer segmentation errors during OCTA generation. To address these challenges, we curate a substantial dataset comprising thousands of OCTA samples with clinical annotations. We develop a 2D classifier based on OCTA projections and assess the impact of segmentation errors qualitatively and quantitatively, proposing two potential solutions. Firstly, we advocate for direct analysis of 3D OCTA volumes using a 2D convolutional neural network trained with additional projection supervision. Secondly, in the absence of 3D data, we propose leveraging cross-instrument OCTA data and employing style transfer techniques, such as CycleGAN, to facilitate domain conversion. To enhance CycleGAN’s performance for classification tasks, we integrate class-related constraints during training. Furthermore, we extend the concept of class-conditioned training to image synthesis by employing diffusion models. Specifically, we fine-tune a latent diffusion model for retinal image generation, incorporating semantic score-based filtering and class-conditioning to maintain the high semantic quality and clinical relevance of generated samples.

Experimental results demonstrate our approach significantly boosts classification accuracy and generalization capabilities, particularly in data-scarce settings, underscoring the potential of CycleGAN- and diffusion-based augmentation methods for robust deep learning classifier training. As a result, our classifiers achieve over 80% accuracy on a four-stage grading task, surpassing human expert accuracy of 60%, irrespective of segmentation errors. Our findings underscore OCTA’s potential as a promising modality for differentiating AMD disease stages and highlight the significance of our contributions in advancing OCTA-based AMD diagnosis.

Chapter 1

Introduction

1.1 Motivation and Objective

The retina, a vital part of the human eye, transforms light into neural signals through a photosensitive layer of optic nerve tissue. At the core of retina lies the macula, essential for detecting color and light intensity, without which detailed and vivid color perception is compromised [7]. Unfortunately, various eye diseases, such as age-related macular degeneration (AMD), choroidal neovascularization (CNV), and diabetic macular edema (DME), can impact the health of the macula. AMD, in particular, is one of the leading causes of severe irreversible vision impairment, contributing to a significant health concern. It is a progressive eye disease associated with abnormal vascular alteration and growth originating from the choroid. Starting from an early non-exudative stage, AMD can progress to an exudative stage where 90% of patients may lose vision [8]. Since the progression of AMD has manifestations associated most commonly with the choroidal neovascular (CNV), early detection of pathological vessels is crucial in optimal treatment management and maintaining vision for AMD patients.

Current retinal imaging technology can detect vessel leakage and fluid, but does not support visualizing the pathogenic vessels in the sub-RPE or subretinal space. For example, color

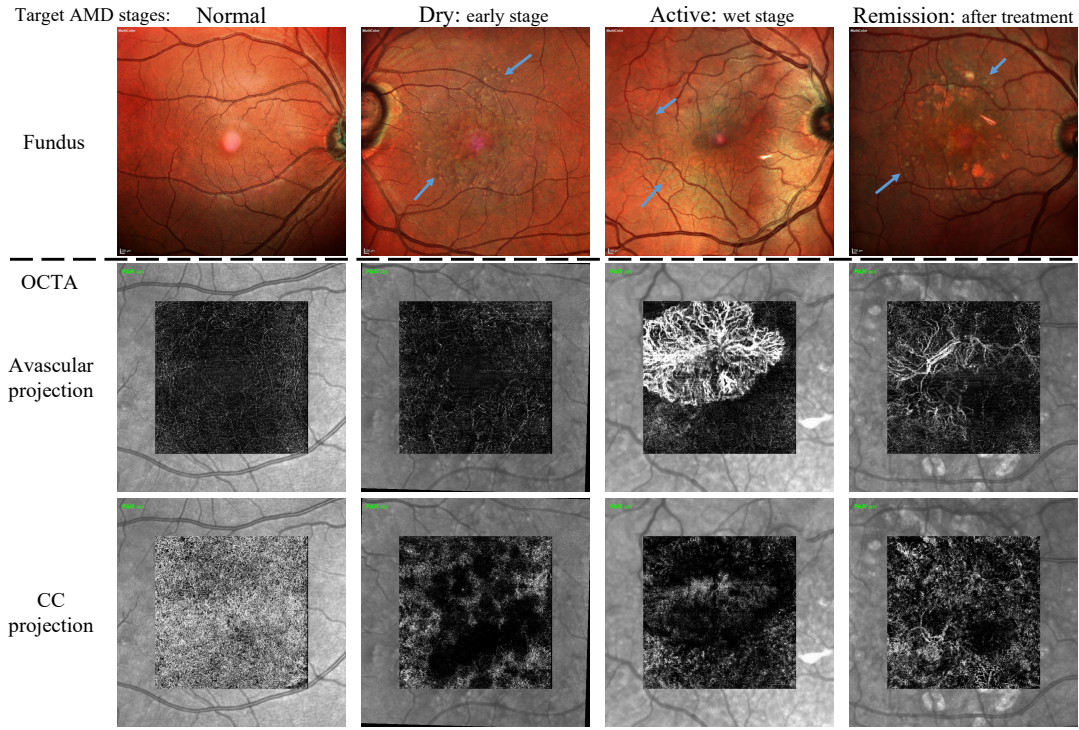


Figure 1.1: Comparison between fundus and OCTA w.r.t. AMD stages. As shown by the blue arrows, all AMD stages exhibit drusens and it is difficult to differentiate each stage based on the pattern of drusens. For instance, in the provided example, the early stage (dry) displays clearer drusens than the progressive stage (active). In contrast, OCTA allows for a distinction between dry and normal stages using the hollows in CC projection, and between active and dry stages with the presence of CNV in avascular projection. It is still an ongoing challenge to tell active stage from remission for human experts, yet this thesis demonstrates it is achievable with the proposed deep classifiers in both 2D and 3D cases.

fundus imaging, the current gold standard, can only reveal large retinal vessels, drusens, and areas of atrophy, which may indicate the presence of AMD, but make it difficult to determine the stage of the disease. Fluorescein Angiography (FA) can show CNV only at a specific time point, which is often short and challenging to capture. Optical Coherence Tomography (OCT) can display retinal layers and fluid but lacks the ability to visualize vessels. In contrast, OCT Angiography (OCTA), as an emerging imaging modality, has the capability to display vascular networks in different retinal layers [9, 10, 11], as depicted in Figure 1.1. It shows superficial and deep vascular complex (SVC and DVC), avascular layer and choriocapillaris (CC). By visualizing the pathological CNV vessels directly, it enables not only an earlier detection, but also a way

to monitor the clinical response to treatment. In Figure 1.1, we provide a comparison between fundus and OCTA w.r.t. different AMD stages.

Unfortunately, even with the above-mentioned benefits, OCTA has not been regarded as the gold standard in clinical decision making yet, because the correlations among visible vessels and response to treatment, prognosis and determining drug activity are not strictly proven. On the clinical side, ophthalmologists are actively searching biomarkers for AMD diagnosis from OCTA, mainly based on manual analysis and their own experience. In this thesis, we present experimental evidence of the informativeness of OCTA from the perspective of data-driven classifiers. We believe that deep learning is capable of this task with two advantages. Firstly, some deep learning algorithms have been proven to surpass human-level performance on natural image classification [12]. Moreover, it is more efficient for computer to handle 3D data or multiple projections than human. Consequently, we expect that deep learning classifiers would identify hidden patterns imperceptible to human eyes and improve AMD diagnosis.

Objective. In this thesis, we focus on OCTA modality *exclusively* and build a series of deep learning based classifiers to explore the effectiveness of OCTA in AMD diagnosis.

1.2 Challenges and Contributions

In our exploration of the potential of OCTA, several challenges have come to light. One of the most prominent is the scarcity of publicly available OCTA datasets. Given the nascent nature of OCTA technology, existing public datasets are typically small in size and unsuitable for developing deep neural networks for AMD analysis. For instance, recent datasets like ROSE [13] and OCTA500 [14] are primarily designed for segmentation purposes and lack crucial AMD stage information. Further details on this issue are discussed in Section 2.3.1. In response to this limitation, we undertook the task of collecting our own dataset (Section 2.3.2), in collaboration with the Jacobs Retina Center at the Shiley Eye Institute, affiliated with UCSD. Additionally in

Chapter 4, we investigated advanced data augmentation techniques, including CycleGAN [15], to enhance the utility of available OCTA data across different imaging instruments. Building on the success of CycleGAN-based dataset expansion, we integrate a state-of-the-art diffusion model [16] to generate additional retinal images, further increasing the size of the training dataset. During this fine-tuning process, particular emphasis is placed on improving the semantic quality of the generated images, as detailed in Chapter 5.

A second challenge arises from the processes involved in generating OCTA projections from raw OCTA 3D data. To visualize vascular networks within distinct retinal layers, OCTA instruments must first segment these layers. However, existing layer segmentation algorithms often fall short in cases of diseased retinas, particularly those affected by AMD, leading to prevalent segmentation errors. Our research delves into the qualitative and quantitative impact of these errors in Section 3.4.1 and proposes two methodologies to mitigate them. Specifically, we advocate for the direct analysis of 3D raw OCTA volumes, as detailed in Section 3.3.3.

Through the resolution of these challenges, our contributions are:

- We curated an extensive dataset consisting of 1478 high-resolution OCTA en-face scans obtained using the Heidelberg Spectralis device, along with 1003 OCTA scans acquired using the Optovue Solix OCTA (Visionix). Each scan is meticulously annotated by retina specialists, ensuring high-quality clinical relevance.
- Our experiments reveal that OCTA projections, conventionally used by ophthalmologists for AMD diagnosis, are highly susceptible to layer segmentation errors. We demonstrate that these segmentation inaccuracies significantly degrade the performance of deep learning-based classifiers.
- To address the impact of segmentation errors, we propose two complementary strategies: 1) direct analysis of raw 3D OCTA volumes to inherently mitigate segmentation-induced artifacts, and 2) augmentation of training data distributions by explicitly including images

affected by segmentation errors, thereby enhancing model robustness.

- In scenarios where 3D raw OCTA volumes are accessible, we adapt a pre-trained 2D convolutional neural network for direct volume classification. By integrating additional projection-based supervision during training, we improve the model’s capability to extract clinically relevant features effectively.
- For cases limited to 2D OCTA projections, we propose training deep classifiers utilizing cross-instrument data augmented via advanced style transfer techniques, specifically CycleGAN. We introduce novel class-specific constraints during CycleGAN training, guided by a pre-trained classifier, to enhance the semantic relevance and classification quality of the generated images, ultimately improving classification accuracy.
- We further explore and validate the feasibility of employing diffusion models for retinal image synthesis. Specifically, we fine-tune a latent diffusion model for OCTA image generation, integrating semantic score-based filtering and class-conditioned modeling to maintain high semantic consistency and clinical accuracy in the generated images. This strategy effectively augments scarce training data and balances class distributions.
- Comprehensive experimental evaluations confirm that our proposed methodologies, combining CycleGAN-based and diffusion-based augmentation techniques, significantly enhance the classification accuracy and generalization capabilities of AMD stage graders. Remarkably, our classifiers achieve an accuracy surpassing 80% in a challenging four-stage AMD grading task, significantly outperforming human experts, who average approximately 60% accuracy, regardless of segmentation errors.
- Collectively, our results underscore the substantial potential of OCTA as a diagnostic modality for differentiating AMD stages and emphasize the practical value of our proposed approaches. By effectively overcoming segmentation-related limitations and data scarcity,

our contributions pave the way for robust, accurate, and clinically impactful OCTA-based AMD diagnosis.

Chapter 2

Preliminaries and Dataset

2.1 Age-related Macular Degeneration

Age-related macular degeneration (AMD) is a predominant cause of irreversible blindness globally, especially affecting individuals aged 55 and older [17]. This degenerative eye disease primarily targets the macula, a crucial region within the central retina responsible for sharp vision. The progressive cell death in the macula significantly impacts central vision, profoundly affecting daily activities and overall quality of life. Please refer to Figure 2.1 for visions with different types of AMD.

The hallmark characteristics of AMD in the retinal exam include pigmentary changes, drusen, choroidal neovascularization, and geographical atrophy. These findings may be present separately or all together, with drusen being often the first hallmark of AMD. Pigment changes manifest as noticeable color variations in the retinal pigment epithelium (RPE) layer, while drusen refers to the formation of yellow deposits in the retina due to the accumulation of waste products. In advanced stages, geographical atrophy emerges, marked by the absence of the RPE layer resulting from the death of choroidal vessels. There are two primary types of AMD: dry and wet (or neovascular) [18]. Dry AMD, affecting approximately 80% of AMD cases, results from

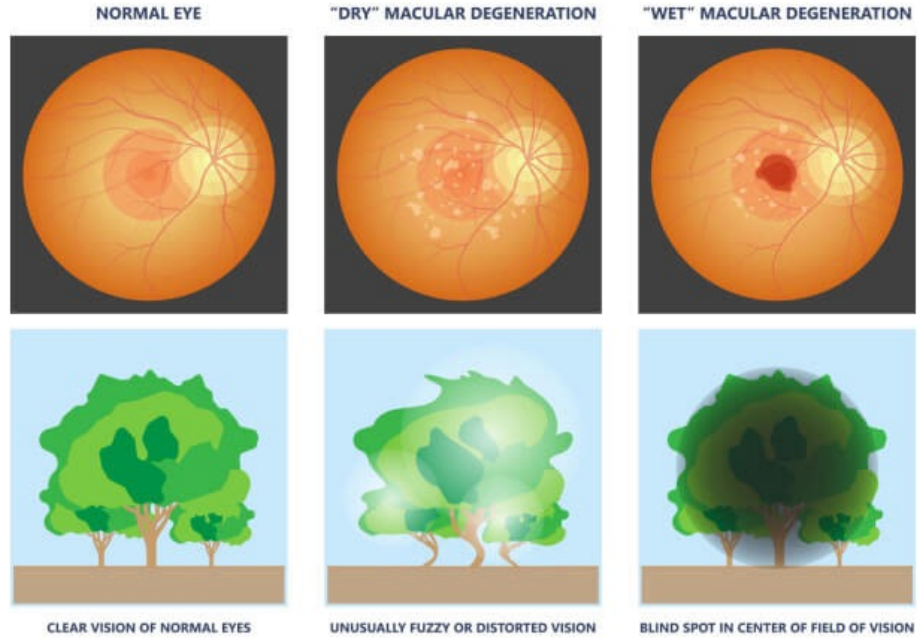


Figure 2.1: Vision with different types of AMD, from All About Vision website [1]. AMD doesn't cause complete blindness, but losing your central vision can make it harder to see faces, read, drive, or do close-up work like cooking or fixing things around the house.

the gradual thinning of the macula with age, accompanied by the development of drusen and pigment abnormalities. In contrast, wet AMD, though less prevalent, poses a more severe threat as it may cause bleeding and rapid central vision loss leading to rapid scarring of the macula if left untreated. The primary distinction lies in the formation of choroidal neovascularization (CNV), a severe complication and defining feature of neovascular AMD (nAMD), which can result in acute and often irreversible vision loss due to active leakage. The leakage of CNV is typically diagnosed through fundoscopy and fluorescein angiography, which have served as the gold standard in clinical practice for decades.

While fluorescein angiography has historically served as the gold standard for detecting CNV in AMD, the emergence of Optical Coherence Tomography Angiography (OCTA) presents a promising alternative. OCTA, despite not yet achieving gold standard status, is noninvasive and offers capability to identify subclinical pathologic changes, making it well-suited for diagnostic imaging and monitoring purposes [19]. OCTA additionally enables the quantification of vascular

metrics that exhibit correlations with clinical disease severity and progression. Specifically, it can be employed for a quantitative assessment of morphological changes in CNV vessels in neovascular AMD in response to anti-VEGF treatment [20]. For more information about OCTA, please refer to Section 2.2.2.

In this thesis, we embark on an exploration of OCTA’s potential in AMD diagnosis utilizing deep learning-based classifiers. By leveraging the capacity of data-driven classifiers to discern hidden patterns and handle 3D data efficiently, we experimentally verify the effectiveness of OCTA in AMD diagnosis. This study represents a comprehensive effort to advance AMD diagnosis using OCTA as the primary imaging modality, with the ultimate goal of enhancing patient outcomes and improving clinical decision-making in the management of this debilitating eye disease.

2.2 Retinal Imaging Modalities

2.2.1 Optical coherence tomography

Optical Coherence Tomography (OCT) is an advanced imaging technique that enables the visualization of cross-sectional details of retinal and choroidal microvasculature. Similar to ultrasound imaging, OCT utilizes light beams instead of sound profiles to capture A-scan signals reflecting from different layers within an object. These signals are then used to create B-scan or cross-sectional images. Unlike ultrasound, OCT does not require direct contact with the eye for signal transmission or reception, and it achieves higher spatial resolution using light. Please refer to Figure 2.2 for an illustration of 3D OCT and OCTA volumes as well as B-scans.

OCT achieves micrometer-level resolution by employing various light sources such as superluminescent diodes, ultrashort-pulsed lasers, or supercontinuum lasers. This method finds numerous applications in imaging anatomical structures, particularly in the medical field where information about internal layers of structures is essential. However, OCT is typically limited to

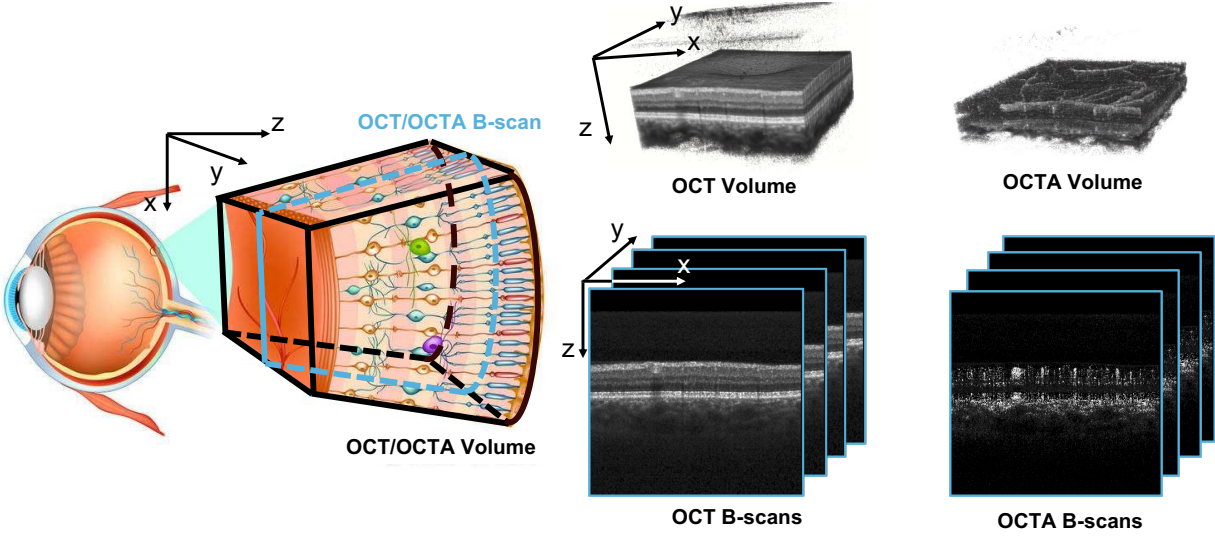


Figure 2.2: 3D view of OCT and OCTA raw volume, modified from Shutterstock [2]. The 3D OCT and OCTA data consist of many B-scans shown as blue box. If we use x, y as spatial direction and z for depth, each B-scan has x and z two dimensions. They are concatenated along y axis to form a 3D volume.

imaging depths of 1-2 mm, beyond which the strength of the scattered beam becomes too weak for effective image reconstruction. There are two main approaches to OCT imaging: Time Domain OCT (TD-OCT) and Frequency Domain OCT (FD-OCT). TD-OCT utilizes interferometry in low-coherence or white light to create cross-sectional images based on differences in reflected light between biological tissue and a reference light beam. On the other hand, FD-OCT eliminates the need for mechanical movement of the reference arm and retrieves depth information through Fourier transform, resulting in faster imaging. Please refer to [21] for detailed introduction of OCT.

Optical Coherence Tomography has revolutionized the diagnosis and management of neovascular Age-related Macular Degeneration (nAMD) by offering high-resolution imaging of intraretinal, subretinal, and subretinal pigment epithelium fluid. This level of detail enables enhanced detection of nAMD activity, crucial for timely intervention and improved visual outcomes. Furthermore, advancements in OCT technology, including improved image analysis algorithms and faster acquisition speeds, have augmented its utility in guiding treatment decisions. Unlike

traditional methods such as fluorescein angiography, which lack three-dimensional anatomical information, OCT provides cross-sectional images of the macula, facilitating precise monitoring of disease progression and response to therapy. While OCT devices are not yet specifically validated for anti-VEGF therapy, their integration into clinical practice has already enhanced treatment decision-making [22]. Additionally, choroidal thickness, as assessed by OCT imaging, serves as a valuable indicator of AMD subtype. Recent studies [23] have elucidated the variable nature of choroidal thickness among AMD patients, revealing differences between wet-type and dry-type AMD. For example, subfoveal choroidal thickness tends to be thinner in eyes with only reticular pseudodrusen compared to those with early AMD or multiple medium or large drusen within the macula [24]. These findings underscore the significance of OCT in providing detailed anatomical information and aiding in the differential diagnosis of AMD subtypes, thereby enabling more precise treatment strategies and improved patient outcomes.

2.2.2 OCT angiography

Ocular diseases often manifest with vascular abnormalities, traditionally captured using invasive dye-based techniques such as fluorescein angiography (FA) and indocyanine green angiography (ICGA). However, these methods require intravenous injection of contrast agents, posing risks of serious side effects. Optical coherence tomography angiography (OCTA) has emerged as a novel, noninvasive alternative for imaging vascular structures without the need for dye [25]. By leveraging the variations in OCT signals caused by flowing red blood cells, OCTA produces high-resolution 3D retinal and vascular angiograms, offering faster, more confident, and repeatable diagnosis compared to dye-based techniques.

The imaging principle of OCTA revolves around the variation of OCT signals in response to blood flow, which is more pronounced than in static tissue. Multiple approaches, including changes in signal intensity and phase, are employed to capture these variations, culminating in the generation of angiographic maps. OCTA offers a 3D dataset representation of vascular

structures within the target tissue, allowing clinicians to navigate through projection images and examine layers in detail. However, the technique necessitates multiple scans of the same location, which can be time-consuming for larger areas. OCTA has shown promise in evaluating common ophthalmologic diseases such as AMD, diabetic retinopathy, artery and vein occlusions, and glaucoma. However, its utility is hindered by differences in hardware and software among OCTA devices from different manufacturers. Variability in imaging results, including vessel visibility and motion artifacts, underscores the importance of acquiring patient images on the same device for accurate follow-up of specific diseases. Moreover, OCTA has limitations in capturing retinal periphery and detecting pathology in those areas. Despite its insensitivity to leakage, limited field of view, and potential for artifacts, OCTA, with proper training and landmark identification, holds promise as a powerful tool for capturing various retinal vascular diseases, inflammatory conditions, tumors, and hereditary dystrophies.

Optical Coherence Tomographic Angiography (OCTA) has emerged as a valuable tool in the detection of choroidal neovascularization (CNV) in wet AMD, owing to its ability to visualize retinal and choroidal circulation with high detail. Numerous studies have demonstrated the effectiveness of OCTA in CNV detection in neovascular AMD, with documented sensitivities ranging from 50% to 87% compared to gold standard fluorescein angiography and structural OCT, alongside specificities of 91% to 98% using commercially available spectral domain devices [26]. Additionally, OCTA facilitates the quantification of vascular metrics correlating with disease severity and progression, allowing for a quantitative assessment of morphological changes in CNV vessels in response to anti-VEGF treatment [20]. This modality offers detailed visualization of neovascular membranes, including structure, size, location, and blood flow, potentially identifying nonexudative “subclinical CNV” lesions before detection with conventional imaging such as OCT and FA. However, despite its benefits, OCTA is not without limitations. Eye motion can introduce noise artifacts, while media opacities and hemorrhage may lead to shadow artifacts. Moreover, normal vasculature can result in projection artifacts, and low-flow neovascular membranes may

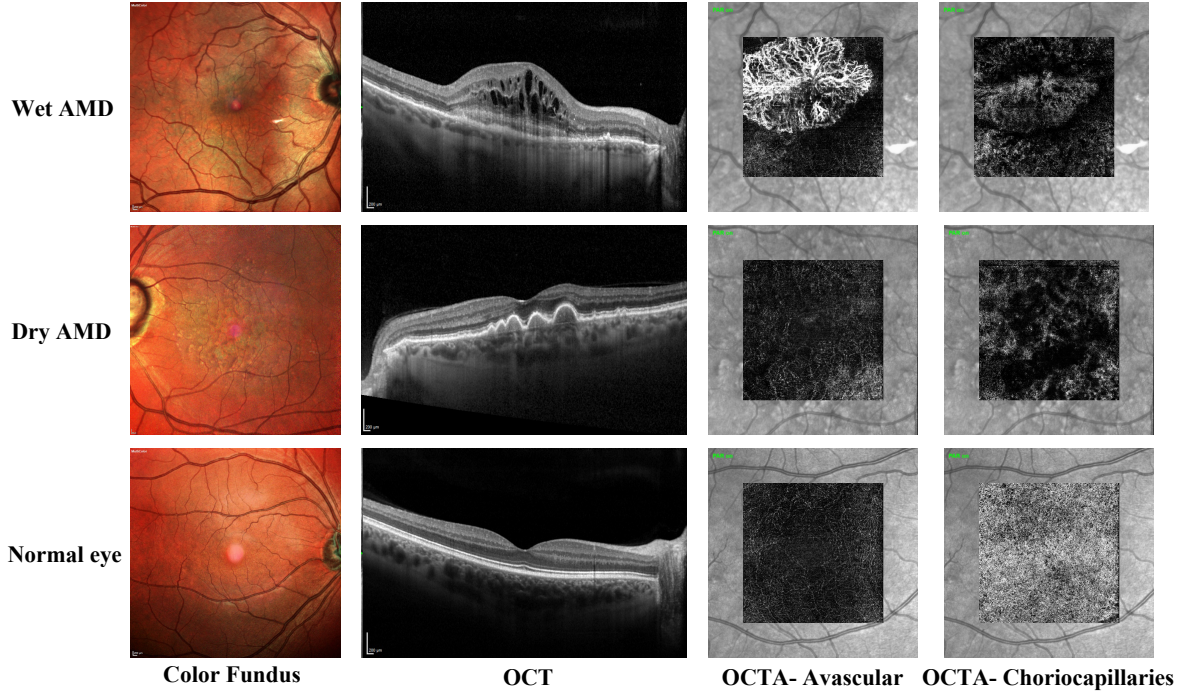


Figure 2.3: Various retinal imaging techniques for diagnosing AMD. In color fundus images, drusen in the macula distinguish AMD from a normal retina. Diagnosing wet AMD is challenging without observing macular hemorrhage. OCT images reveal fluid in wet AMD, drusen in dry AMD, and no abnormalities in normal eyes. In OCTA, wet AMD reveals pathological CNV in the avascular layer, and dry AMD shows halos in the CC layer. OCTA can diagnose quiescent CNV, indicating early wet AMD before leakage occurs.

not be adequately visualized due to resolution limitations. Nonetheless, ongoing advancements in algorithms and scanning technologies hold promise for overcoming these limitations and enhancing the precision and resolution of OCTA images [27].

In Figure 2.3, we provide some examples to show how different retinal imaging techniques help in AMD diagnosis. For color fundus image, it is easy to distinguish retina with AMD from normal using the drusens shown in the macula area, but it is hard to diagnose wet AMD unless a macular hemorrhage is found, which is a late symptom of AMD. Using OCT, we can see find fluid for wet AMD, some bumps (drusens) for dry AMD and nothing abnormal for normal eyes. Using OCTA, we can directly see pathological CNV vessels in the avascular layer for wet AMD and the hallows in Choriocapillaries (CC) layer indicate the exist of drusens meaning dry AMD.

In normal retina, there would be nothing in the avascular layer and noise in the CC layer since the vessel within this layer is too thin to capture their pattern with our current instruments.

In Figure 2.3, we illustrate how various retinal imaging techniques aid in diagnosing AMD. In color fundus images, distinguishing AMD from a normal retina is relatively straightforward due to the presence of drusen in the macula area. However, diagnosing wet AMD is challenging unless a macular hemorrhage, a late-stage symptom, is observed. OCT images reveal fluid accumulation in wet AMD, drusen in dry AMD, and no abnormalities in normal eyes. For OCTA in a normal retina, the avascular layer shows no vessels, and the choriocapillaris (CC) layer appears noisy due to the thinness of the vessels, which current instruments cannot capture clearly. In contrast, for wet AMD, OCTA reveals pathological CNV in the avascular layer, while halos in the CC layer indicate the presence of drusen in dry AMD. The capability of OCTA to directly visualize pathological CNV vessels allows for the diagnosis of a new stage of AMD, termed quiescent CNV, where pathological vessels grow without leakage. This provides an early signal of wet AMD before it becomes exudative, facilitating earlier monitoring and detection.

While earlier detection through OCTA may lead to improved monitoring and treatment initiation for high-risk patients, the impact on visual prognosis remains to be fully elucidated. In this thesis, we embark on an exploration of the effectiveness of OCTA in the detection and characterization of CNV in AMD.

2.3 OCTA Dataset

2.3.1 Existing OCTA datasets

The advancement in deep learning has led to significant progress in the field of retinal disease diagnosis and management. Various challenges have been organized to evaluate the performance of computer-aided diagnosis systems on different retinal diseases, such as glaucoma and AMD. The GAMMA challenge [28] is one such challenge, which provides 2D fundus image

Table 2.1: Comparison of public OCTA datasets with the dataset collected by us. Please note that our dataset is specifically curated for AMD stage grading and has the largest number of total and AMD samples available.

OCTA dataset	Task	2D/3D	# of samples	# of AMD
PREVENT	Segmentation	2D only	55	Unknown
ROSE [13]	Segmentation	2D only	229	Unknown
OCTAGON [30]	Segmentation	2D only	213	0
FOCTAIR [30]	Registration	2D only	86	Unknown
OCTA-500 [14]	Segmentation	3D and 2D	500	49
Ours	Classification	3D and 2D	2481	2047

and 3D OCT volume, focusing on glaucoma diagnosis. The ADAM challenge [29] evaluates the performance of automated AMD diagnosis based on fundus image. Although these challenges have provided valuable insights into the development of automated diagnosis systems, they do not include OCTA information in their datasets, which is the key objective of this thesis. Consequently, it is impossible for us to experiment on those datasets.

In this thesis, an exhaustive examination of existing OCTA datasets was undertaken to unveil their limitations in facilitating AMD stage grading, emphasizing the pressing need for a dataset meticulously tailored for this purpose. Table 2.1 provides a comprehensive overview of currently available OCTA datasets.

Notably, datasets such as PREVENT¹, ROSE [13], and OCTAGON [30] primarily concentrate on providing OCTA retinal scans with annotations for vascular segmentation, albeit with limited sample sizes and AMD annotations. In detail, the PREVENT dataset comprises OCTA retinal scans from 11 participants, focusing on parafoveal OCTA images of the superficial layer obtained with the RTVue-XR Avanti OCT system. The dataset includes associated manual segmentations of the vasculature, with each participant contributing five subimages from distinct regions of interest, resulting in a total of 55 high-quality retinal images for analysis. ROSE [13] comprises two subsets: ROSE-1 and ROSE-2. ROSE-1 includes 117 OCTA images from 39 subjects, with annotations of superficial, deep, and inner retinal vascular plexus, captured using

¹<https://datashare.ed.ac.uk/handle/10283/3528>

the RTVue XR Avanti SD-OCT system. ROSE-2 consists of 112 OCT-A images from 112 eyes, focusing on superficial vascular complex (SVC) annotations, obtained with the Heidelberg OCT2 system. The OCTAGON dataset [30] comprises 144 healthy and 69 diabetic Angiography, categorized into four groups: 3x3 superficial, 3x3 deep, 6x6 superficial, and 6x6 deep. Each group includes approximately 36 healthy and 17 diabetic OCTA images. The recently introduced OCTA-500 dataset stands out for its comprehensive nature, encompassing images from 500 subjects with extensive annotations, including segmentation labels, offering a promising avenue for OCTA research. However, despite its breadth, the scarcity of AMD samples in existing datasets renders them inadequate for OCTA-based AMD classification. Besides segmentation, there is one dataset targeted on registration task: The FOCTAIR dataset [30] comprises 86 cases from 29 patients, each representing a patient’s eye. It includes three images per case: one Fluorescein Angiography (FA) image and two OCTA images of the superficial capillary plexus, corresponding to two different zoom levels (3x3mm and 6x6mm), totaling 172 possible registrations. However, it lacks the requisite AMD annotations for classification purposes either. Thus, the scarcity of suitable datasets underscores the significance of the dataset collected in this study.

To overcome these limitations and advance OCTA-based AMD diagnosis research, we have meticulously curated a well-organized OCTA dataset specifically tailored for AMD stage grading. This dataset boasts the largest number of total and AMD samples available, thereby addressing the intricate requirements of AMD stage classification. Detailed insights into the collection process and dataset statistics will be provided in the subsequent section.

2.3.2 Our OCTA dataset

The dataset collected in this study was obtained through a cross-sectional retrospective analysis conducted at the Jacobs Retina Center, Shiley Eye Institute, University of California San Diego, California, USA. It encompasses a total of 1478 high-resolution OCTA en-face scans and raw volumes obtained using the Heidelberg Spectralis device from patients treated

at the Jacobs Retina Center between 2021 and 2023. Detailed clinical data, including imaging reports and electronic medical records managed through EPIC software, were meticulously reviewed to extract relevant information such as diagnosis and details of anti-vascular endothelial growth factor (anti-VEGF) therapy. Additionally, to augment our dataset, we gathered 1003 high-resolution 6x6 mm OCTA en-face scans from the Optovue Solix OCTA (Visionix) device at the Medical University Gdańsk, Poland. This dual-source approach provides a robust foundation for advancing AMD research and developing novel diagnostic methodologies using OCTA.

Dataset collection

Data Acquisition: For the Heidelberg subset, OCT and OCTA scans centered at the macula were conducted using the Heidelberg Spectralis HRA + OCT device (Heidelberg Engineering, Heidelberg, Germany). The OCTA scan parameters were standardized to include 512 B-scans, with a pattern size of $3 \times 3 \text{ mm}/10^\circ \times 10^\circ$, a 6 mm interval between B-scans, and an average of 5 ART images. Each 3D raw OCTA sample consists of 512 B-scans in the y direction, with 512 pixels in the x direction and 496 pixels in the z direction. For 2D en-face OCTA projection, each sample contains 12 projections within different retinal layers, with our investigation focusing on four specific layers: superficial vascular complex, deep vascular complex, avascular complex, and choriocapillaris. Similar acquisition methods were employed for the data collected at the Medical University of Gdansk, Poland, using the Optovue Solix camera. Notably, the field of view of scans obtained with Optovue Solix was larger compared to Heidelberg. Patients underwent OCT and OCTA scans with Optovue Solix using parameters including 512-line B-scans and a scan size of $6 \times 6 \text{ mm}/20^\circ \times 20^\circ$. Each exported OCTA scan comprised four en-face projections: Angio-Superficial, Angio-Deep, Angio-Outer Retina, and Angio-Choriocapillaris, providing comprehensive visualization of vascular structures. Refer to Figure 2.4 for a visual comparison of vascular en-face 2D projections captured using both Heidelberg Spectralis and Optovue Solix instruments.

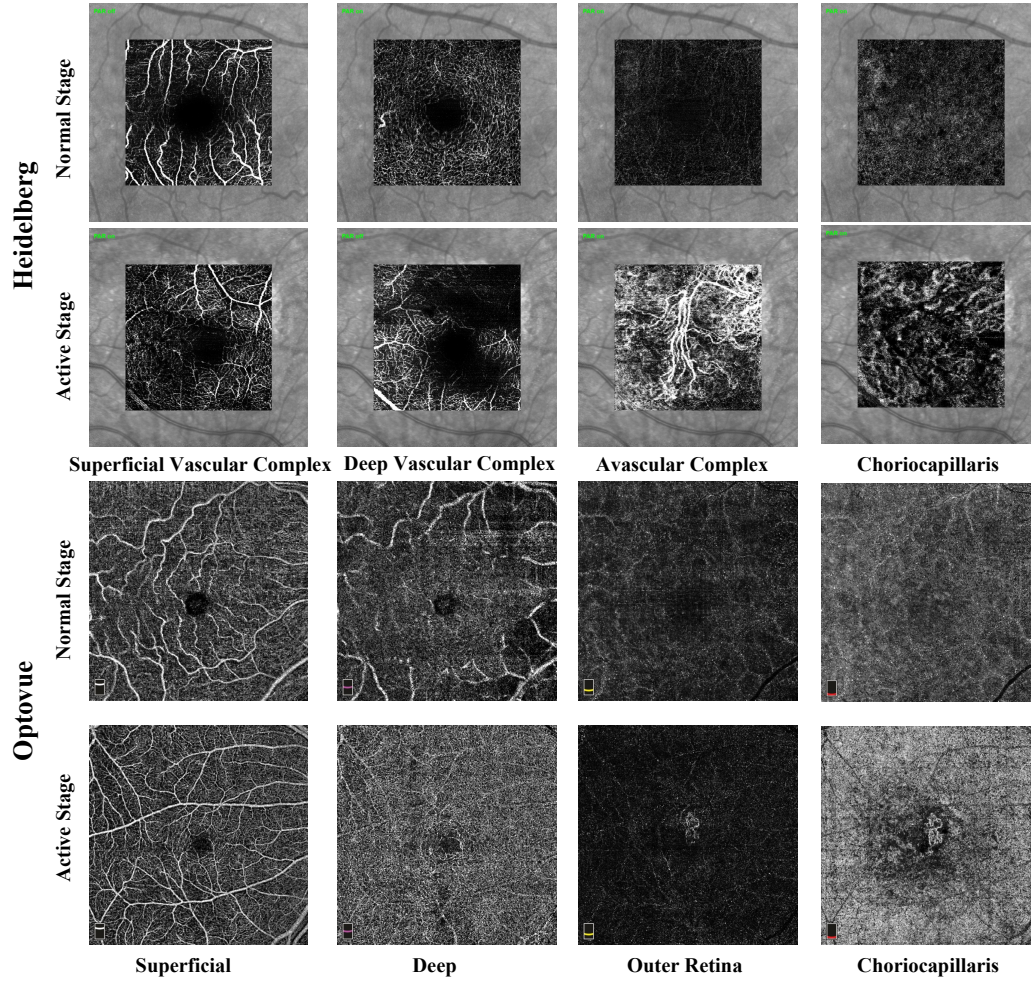


Figure 2.4: Examples of OCTA images in active stage (group1) and normal stage (group4) of AMD captured using Heidelberg and Optovue instruments respectively. The images in active stage illustrate CNV in the avascular slab for Heidelberg and in the outer retina and choriocapillaris slab for Optovue.

Data selection. Inclusion criteria for both the Heidelberg and Optovue datasets were stringent, requiring patients to be aged over 50 with a confirmed diagnosis of AMD and providing high-quality OCTA scans devoid of motion artifacts or background noise. This entailed a quality index (Q) exceeding or equal to 25 for images acquired with Heidelberg Spectralis and a Scan Quality score exceeding or equal to 7 (on a scale from 0 to 10) for scans obtained with Optovue Solix OCTA. Additionally, healthy, age-matched subjects were included as normal controls in both datasets. Conversely, scans with CNV of etiologies other than AMD, along with those

exhibiting motion artifacts, incompleteness, background noise, or poor quality due to media opacity (with a quality index $Q < 25$ in Heidelberg Spectralis and score < 7 in Optovue Solix), were excluded from the analysis to ensure the reliability and accuracy of the dataset. Inclusion criteria applied at the individual eye level.

Data labeling. Clinical activity was assessed by two experienced retina specialists based on the presence of subretinal and/or intraretinal fluid detected in OCT B-scans acquired on the same day as the OCTA scan. This evaluation relied on information gleaned from electronic chart reviews, encompassing diagnosis and anti-VEGF treatment records, without direct examination of the OCTA images, establishing it as the “gold standard” for labeling. Any discrepancies in labeling were meticulously resolved through review by a senior retina specialist, who rendered the final diagnosis following a comprehensive assessment of the imaging results. Patients were categorized into four distinct groups based on their clinical activity: Group 1, designated as “wet active AMD”, included individuals presenting with CNV and detectable fluid in B-scan OCT scans, comprising newly diagnosed patients, treatment-naïve individuals, and those undergoing anti-VEGF therapy with observable fluid on OCT. Group 2, labeled “wet AMD in remission”, comprised patients with treated CNV but devoid of fluid on OCT scans. Group 3 consisted of cases of dry AMD lacking CNV or fluid on OCT scans, while Group 4 encompassed healthy eyes devoid of AMD or any other retinal pathology. Each eye was evaluated independently.

Segmentation error analysis. In addition to the annotation of OCTA data, we performed a comprehensive examination of layer segmentation for the Heidelberg dataset. The layer segmentation was initially conducted using the automated segmentation feature embedded within the Heidelberg software. Subsequently, this automated segmentation underwent rigorous evaluation by two retina specialists responsible for data labeling. To ensure accuracy, each OCTA scan was inspected for segmentation errors directly on the Heidelberg SPECTRALIS machine using the Heyex-2 software, systematically scanning through the 512-line B-scans. Specifically, the integrity of the Bruch membrane line was meticulously assessed, while inner retinal segmentation

Table 2.2: Data statistics of our Heidelberg and Optovue datasets with clinical explanation.

Label	Clinical explanation	Heidelberg		Optovue	
		Amount	Total	Amount	Total
1(Active)	new active	75	496	109	263
	active treated	421		154	
2(Remission)	remission ongoing treatment	230	383	118	212
	midterm remission	69		30	
	long term remission	84		64	
3(Dry)	high risk CNV fellow	92	409	105	284
	high risk no CNV fellow	103		46	
	low risk drusen	214		133	
4(Normal)	Normal	190	190	244	244

was not reviewed. Any identified segmentation errors, whether major or minor, pertaining to the Bruch membrane were promptly rectified and meticulously annotated as ‘Y’ for major errors or ‘M’ for minor errors. Instances where segmentation was deemed accurate and exhibited precise delineation of the retinal layers were denoted as ‘N’ for no error. This rigorous analysis benefits subsequent analysis and interpretation presented in Section 3.3.2.

Dataset statistics

Table 2.2 categorizes various stages of AMD along with their clinical explanations and the number of samples collected from two imaging devices, Heidelberg and Optovue.

- **Active:** This category includes “new active” cases where OCTA was done on newly diagnosed wet AMD or without any prior treatment, and “active treated” cases with ongoing treatment and fluid present on OCT. The Heidelberg dataset contains 75 new active and 421 active treated samples, totaling 496. The Optovue dataset includes 109 new active and 154 active treated samples, totaling 263.
- **Remission:** This includes “remission ongoing treatment” where OCTA was done within months of the last treatment or during ongoing treatment with no fluid on OCT, “midterm remission” (more than 3 months but less than a year since the last treatment with no fluid),

and “long term remission” (over a year since the last treatment with no fluid). Heidelberg has 230 remission ongoing treatment, 69 midterm remission, and 84 long term remission samples, totaling 383. Optovue has 118 remission ongoing treatment, 30 midterm remission, and 64 long term remission samples, totaling 212.

- **Dry:** This category includes “high risk CNV fellow” with no fluid on OCT but with confluent drusen in the scanned eye and CNV in the fellow eye, “high risk no CNV fellow” with no fluid on OCT and confluent drusen in the scanned eye but no CNV in the fellow eye, and “low risk drusen” indicating early-intermediate AMD with no fluid and few drusen. Heidelberg has 92 high risk CNV fellow, 103 high risk no CNV fellow, and 214 low risk drusen samples, totaling 409. Optovue has 105 high risk CNV fellow, 46 high risk no CNV fellow, and 133 low risk drusen samples, totaling 284.
- **Normal:** This category includes samples with no retinal pathology. Heidelberg has 190 normal samples, and Optovue has 244 normal samples, each totaling their respective amounts.

This table clearly illustrates the distribution of AMD stages and their clinical conditions across the two imaging devices, emphasizing that the OCTA dataset collected in this study is specifically tailored for AMD stage grading, featuring the largest data amount and detailed clinical annotations.

Chapter 2, in part, is a reprint of the material as it appears in 2023 International Conference on Computer Vision (ICCV) Workshop [3], Haochen Zhang, Anna Heinke, Carlo Miguel B. Galang, Daniel N. Deussen, Bo Wen, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Cheolhong An. “Robust AMD Stage Grading with Exclusively OCTA Modality Leveraging 3D Volume”, in *ICCV* workshop, 2023. The dissertation author is the primary author of this paper. It, in part, is also a reprint of the material as it appears in *RETINA* [4], Anna Heinke, Haochen Zhang, Daniel N. Deussen, Carlo Miguel B. Galang, Alexandra Warter, Fritz G.

P. Kalaw, Dirk-Uwe G. Bartsch, Lingyun Cheng, Cheolhong An, Truong Q. Nguyen, and William R. Freeman. “Artificial Intelligence for OCTA-based Disease Activity Prediction in Age-Related Macular Degeneration”, *RETINA*, 2022. The dissertation author is a co-primary contributor to this work responsible for AI classifier development and experiment results analysis.

Chapter 3

Robust AMD Stage Grading with 3D OCTA Volume

In this chapter, we present our first step to build deep learning based AMD stage graders focusing on OCTA modality only.

3.1 Introduction

Age-related Macular Degeneration (AMD), one of the leading causes of severe irreversible vision impairment, is a progressive eye disease associated with abnormal vascular alteration and growth originating from the choroid. Starting from an early non-exudative stage, AMD can progress to an exudative stage where 90% of patients may lose vision [8]. Since the progression of AMD has manifestations associated most commonly with the choroidal neovascular (CNV), early detection of pathological vessels is crucial in optimal treatment management and maintaining vision for AMD patients.

However, imaging vessels within different retina layers is not supported by typical retinal imaging techniques. For example, fundus imaging can only reveal large retinal vessels, drusens,

and areas of atrophy, which may indicate the presence of AMD, but make it difficult to determine the stage of the disease. Fluorescein Angiography (FA) can show CNV only at a specific time point, which is often short and challenging to capture. Optical Coherence Tomography (OCT) can display retinal layers and fluid but lacks the ability to visualize vessels. In contrast, OCT Angiography (OCTA), as an emerging imaging modality, has the capability to display vascular networks in different retinal layers [9, 10, 11], as depicted in Fig. 1.1 and Fig. 3.2. It shows superficial and deep vascular complex (SVC and DVC), avascular layer and choriocapillaris (CC). By visualizing the pathological CNV vessels directly, it enables not only an earlier detection, but also a way to monitor the clinical response to treatment. In Fig. 1.1, we provide a comparison between fundus and OCTA w.r.t. different AMD stages.

Unfortunately, even with the above-mentioned benefits, OCTA has not been regarded as the gold standard in clinical decision making yet, because the correlation between vessels in OCTA and AMD stages is not strictly proven. On the clinical side, ophthalmologists are actively searching biomarkers for AMD diagnosis from OCTA, mainly based on manual analysis and their own experience. In this work, we present experimental evidence of the informativeness of OCTA from the perspective of data-driven classifiers. We believe that deep learning is capable of this task with two advantages. Firstly, some deep learning algorithms have been proven to surpass human-level performance on natural image classification [12]. Moreover, it is more efficient for computer to handle 3D data or multiple projections than human. Consequently, we expect that deep learning classifiers would identify hidden patterns imperceptible to human eyes and improve AMD diagnosis.

In this chapter, we focus on OCTA modality only and build a series of deep learning based AMD stage graders. We summarize our contributions as follows:

- We experimentally verified that the OCTA projections, which ophthalmologists usually use for diagnosis, are easily affected by layer segmentation errors. Those errors degrade the classification performance.

- We propose to use 3D raw OCTA volume to avoid the impacts of those errors. To achieve this, we modify a pretrained 2D network to perform volume classification. We also adopt an additional projection supervision to facilitate training of shallow feature extractor.
- Experimental results show that the proposed classifier can achieve the accuracy of more than 80%, regardless of the presence of layer segmentation errors. These results prove the effectiveness of our methods and suggest that OCTA is a promising modality to distinguish various stages of AMD disease.

3.2 Related Works

3.2.1 OCTA analysis in computer vision

In recent years, OCTA has emerged as a valuable tool in ophthalmology, offering a non-invasive way to visualize and analyze the vascular network of the retina. Consequently, most OCTA-based works in the realm of computer vision have focused on segmentation tasks. Alam et al. [31] used U-Net for artery-vein classification and adopted transfer learning to compensate for the small dataset. In [32], a multi-scaled encoder-decoder neural network was employed to detect the avascular area in OCTA projections. Li et al. [33] proposed an Image Magnification Network to segment the thin-and-small capillary plexus using an up-sampling encoding and down-sampling decoding U-Net. Additionally, Li et al. [34] developed a method to segment vessels with 3D OCTA inputs, eliminating the need for projection images and retinal layer segmentation.

In addition to segmentation, some deep learning-based OCTA classification works have emerged. For example, Le et al. [35] utilized the VGG16 network to classify diabetic retinopathy stages: diabetic with and without retinopathy, and healthy subjects. They pre-trained the classifier on ImageNet [36] and fine-tuned the last seven convolution layers on OCTA data. Lin et al. [37]

advanced this approach by performing classification and segmentation simultaneously, using boundary shape and distance maps as additional supervision to improve accuracy. Beyond classification and segmentation, other researchers have focused on 3D vessel reconstruction [38], projection quality assessments [39], and improving en face OCTA generation [40]. Although these works have shown promising results, none of them have addressed the grading of AMD stage, which is a critical task in the clinical management of AMD patients.

3.2.2 AMD diagnosis with deep learning

To the best of our knowledge, there is no existing work on AMD diagnosis using only the OCTA modality. Instead, previous studies predominantly rely on color fundus photography, FA, and more recently, OCT modalities. Alqudah et al. [41] trained a customized CNN to classify retinal images into five distinct stages of AMD based on OCT B-scans. Similarly, Motozawa et al. [42] first classified images as AMD or non-AMD and then identified the presence of exudative changes. Das et al. [43] employed a fusion of multi-scale deep image features to classify OCT images into categories such as diabetic macular edema (DME), drusen, CNV, or normal. Fang et al. [44] proposed a lesion-aware CNN with a lesion detection network, simulating ophthalmologists' focus on lesion-related regions to enhance classification accuracy. He et al. [45] used a label smoothing GAN to generate synthetic OCT images, identifying drusen, CNV, DME, and normal eyes. This approach expanded the dataset, addressing overfitting issues associated with limited OCT training samples. Our study aims to fill the gap in the literature by exclusively utilizing the OCTA modality for AMD stage grading, leveraging advanced deep learning techniques.

In addition to stage classification, Banerjee et al. [46] combined hand-crafted and CNN features in an LSTM to predict AMD progression. Motozawa et al. [42] first classified images as AMD/no AMD and then identified the presence of exudative changes. Rakocz et al. [47] designed SLIVER-net to classify risk factors for AMD progression, which could operate on both 2D B-scans and 3D volumes. Russakoff et al. [48] predicted the likelihood of conversion

from early/intermediate to advanced AMD. Furthermore, several recent works [49, 50, 51] have employed multimodal images, such as fundus photographs, OCT B-scans, and OCTA projections, to grade AMD. Specifically, Vaghefi et al. [49] used a multi-input CNN analyzing OCT, OCTA, and color fundus photographs to increase the accuracy of diagnosing intermediate dry AMD. Similarly, Jin et al. [50] utilized OCT and OCTA to assess CNV in AMD. In this chapter, we focus on the latest work by Thakoor et al. [51] for comparison, which utilizes OCT B-scans, OCT projections, and OCTA projections with a customized 3D CNN.

3.2.3 OCTA datasets

The advancement in deep learning has led to significant progress in the field of retinal disease diagnosis and management. Various challenges have been organized to evaluate the performance of computer-aided diagnosis systems on different retinal diseases, such as glaucoma and AMD. The GAMMA challenge [28] is one such challenge, which provides 2D fundus image and 3D OCT volume, focusing on glaucoma diagnosis. The ADAM challenge [29] evaluates the performance of automated AMD diagnosis based on fundus image. Although these challenges have provided valuable insights into the development of automated diagnosis systems, they do not include OCTA information in their datasets, which is the key objective of this chapter. Consequently, it is impossible for us to experiment on those datasets. In the Chapter 2.3.1, we report the detailed information about existing OCTA datasets to show their limitation in OCTA based AMD stage grading. In this chapter, we experiment with an OCTA dataset collected by ourselves, which has the largest number of AMD samples available and is specifically curated for AMD stage grading task.

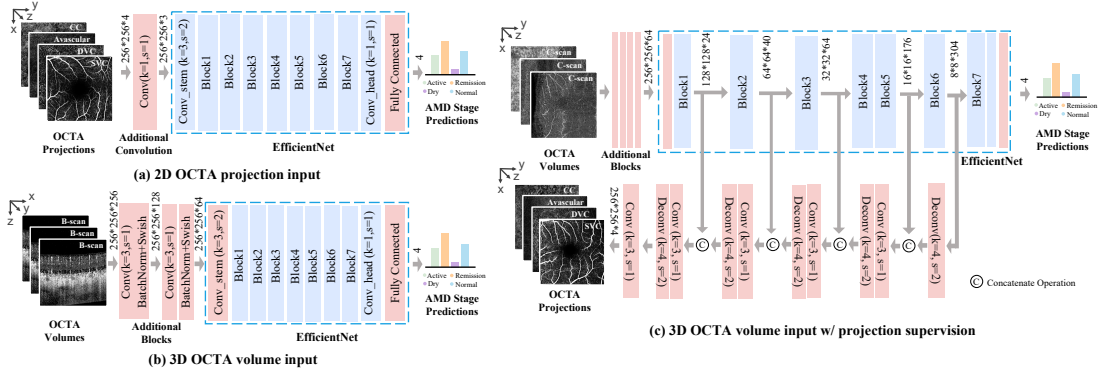


Figure 3.1: The proposed network structures for (a) 2D projections, (b) 3D volumes and (c) 3D volumes with 2D projection supervision. The layers in blue have pretrained weights while those in red are trained from scratch.

3.3 Proposed Methods

3.3.1 2D classifier based on OCTA projection images

In clinical practice, ophthalmologists usually refer to OCTA projections for diagnosis, inspiring related classifiers [35, 51] using the same inputs. In this section, we also develop a baseline classifier with 2D OCTA projection inputs, for analysis and comparison.

Classifier structure. Different from existing method [51], which used a custom CNN without pretraining, our approach utilizes a well established image classification network as backbone. Moreover, we pretrain the backbone with ImageNet [36], and subsequently fine-tune it with our OCTA projections. As shown in Figure 3.1 (a), we adopt the EfficientNet in our network, because it is reported to achieve the best trade-off between performance and model size [52]. Since we set up four input channels to take four OCTA projections, we include an additional convolution layer with kernel size 1 before the EfficientNet to address channel mismatching. Additionally, since we only have four target categories, we adjusted the output of the last fully-connected layer to match the number of categories.

Warmup strategy. Consequently as shown in Figure 3.1, the layers are divided into two groups: the red layers with no pretrained weights and the blue layers pretrained with Imagenet.

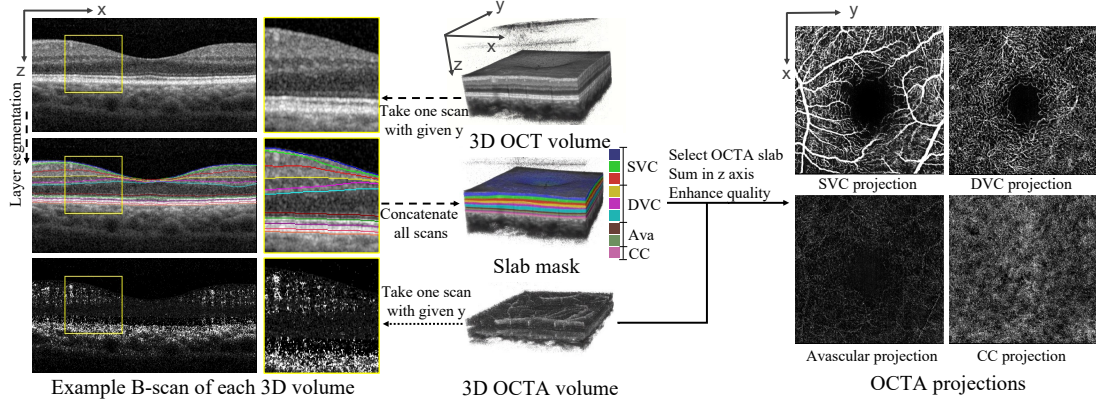


Figure 3.2: Illustration of the interrelationships among OCT and OCTA raw volume, B-scans, and OCTA projection. A single B-scan is a cross-sectional slice of the 3D volume with a specific y-axis value. Retinal slab masks are derived from retinal layer segmentation in each B-scan of 3D OCT volume. OCTA projections are generated by summing up the motion responses in selected OCTA slabs followed by quality enhancement.

Since different groups have different initialization weights, the red layers could disrupt the tuning of the blue ones if fine-tuned all the layers together. So we use a warmup strategy as follows. We first freeze all the blue layers and train only the red ones for 600 epochs. During this step, we also train all the BatchNorm layers to better transfer from natural images distribution to OCTA projections distribution. Then we finetune all the layers together for another 900 epochs with a smaller learning rate.

3.3.2 Presence of layer segmentation errors

During the development of our 2D classifier, we find that OCTA projections are not always reliable due to their sensitivity to the quality of retinal layer segmentation, which plays an important role in OCTA projection generation. This problem is common but often overlooked in most published literature [35, 51]. It is worth noting that previous research [34] has also reported that failures in layer segmentation can lead to difficulties in OCTA vessel segmentation. In this section, we aim to investigate the prevalence of layer segmentation errors and their impact in context of AMD stages grading.

Table 3.1: Distribution of error-free and error-prone samples and associated error rates. Clearly, samples with more severe AMD have larger error rate. The overall error rate shows layer segmentation error is a common problem in OCTA projections. Please refer to Figure 3.3 for visual indication of the detrimental effects of segmentation errors on the quality of the OCTA projections.

sample type	Active	Remission	Dry	Normal	Total
# w/ seg. error	138	91	57	2	288
# w/o seg. error	52	39	90	61	242
error percentage	72.6%	70%	38.8%	3.2%	54.3%

OCTA projection generation. Raw OCTA data capture the movements of blood in a 3D retinal space which are difficult to interpret by humans. Therefore, OCTA imaging machines commonly project raw OCTA volumes onto 2D images to enhance their visual interpretation. The projection process may differ among commercial instruments. Here, we consider the image taken by Heidelberg¹ as an example [53]. As illustrated in Figure 3.2, the Heidelberg software estimates the boundaries of different retinal layers to divide the 3D space into several slabs. Within selected slabs, which are determined by anatomical criteria, it calculates the summation of OCTA responses along z-axis to generate a 2D image. Additionally, the software employs a contrast function and a projection artifact removal algorithm to enhance the image quality. When executed successfully, these steps produce highly informative and visually appealing 2D images that are easily interpretable by doctors.

Influence of segmentation errors. Unfortunately, the estimated layer boundaries in the first step are not always accurate, resulting in segmentation errors that significantly impact the quality of OCTA projections. Since most commercial instruments usually estimate those boundaries based on image gradient and graphcut algorithm [54], which is not robust enough, the layer segmentation errors are actually prevailing, especially for distorted retina with AMD disease. To gain a better understanding of the magnitude of the problem, we conduct a manual check of 530 OCTA samples from different AMD stages and report the results in Table 3.1. Not surprisingly, we find that almost three-fourths of samples in the active stage have layer

¹The Heidelberg HRA+OCT Spectralis System, version 1.11.2.0 (Heidelberg Engineering, Heidelberg, Germany)

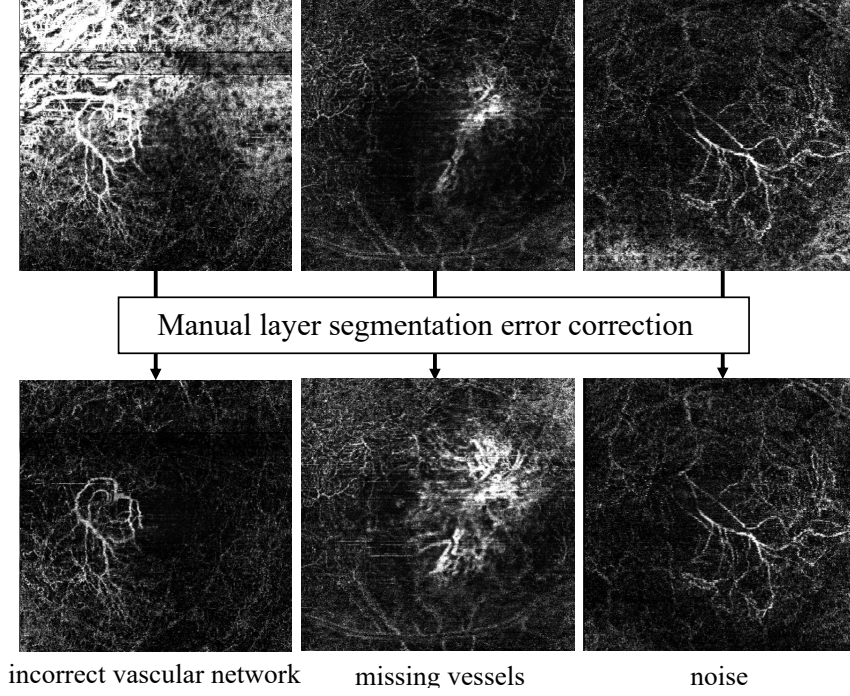


Figure 3.3: Examples of avascular projection w/o and w/ manual layer segmentation error correction by human experts. Layer segmentation errors lead to incorrect vascular networks, missing vessels and noise in OCTA projections, which complicates the classification for both ophthalmologists and neural networks.

segmentation errors. The overall error rate among 530 samples is as high as 54.3% and, more accurately, we can calculate the balanced overall error rate by averaging the last row of Table 3.1, which is 46.2%. These findings indicate that the problem of layer segmentation error is pervasive and requires urgent attention. As shown in Figure 3.3, layer segmentation errors lead to incorrect vascular networks or missing vessels in OCTA projections, which complicates the classification for both ophthalmologists and neural networks. The influence of layer segmentation errors on deep classifier is quantified and discussed in Section 3.4.1.

3.3.3 Avoid segmentation errors with 3D input

Since the errors in layer segmentation significantly affect the quality of OCTA projections, we propose to directly apply raw OCTA volume² for classification. In this section, we provide a detailed description of our method, which utilizes a 2D convolutional neural network to analyze 3D OCTA data. We then delve into the reasons behind our choice of channel dimension and how we further improve the training process to achieve optimal performance.

2D backbone for volume classification. Considering that there is no available large-scale 3D dataset for pretraining a 3D classifier, we use a 2D network with pretrained weights to analyze 3D data. It means that we take one dimension of 3D as channel and the other two as spatial. As shown in Figure 3.1 (b), we gradually reduce the input channel by extending the additional convolution to two Conv-BN-Swish blocks with kernel size 3. Each block divides the channels by 2 and the input channel of the EfficientNet is ultimately revised to a desired number, i.e. 64 in our experiments. Based on the ablation experiments reported in Table 3.5, we find that better accuracy is achieved by treating the dimensions of B-scan as spatial and incorporating different B-scans in the channel dimension, i.e. taking the y-axis in Figure 3.2 as the channel.

3D input pre-processing. In addition to the network structure, we apply two important pre-processing steps for the input 3D OCTA volume. As shown in Figure 3.2, since the informative parts of 3D OCTA volume is thinner than its OCT counterparts, directly using the whole volume is inefficient and may lose information when downsampling is executed. Thus, we propose to crop the volume with layer segmentation. Although segmentation errors are prevailing, the top boundary is usually reliable. To keep the physical resolution, we find the maximum thickness through all samples and then crop from the top boundary until achieving the maximum thickness with some margins. The second important part is data augmentation where we find modifying 2D crop-and-resize augmentation to 3D version helps the training significantly.

²OCTA volume here represents for the raw blood motion responses in 3D space before projection. No structural OCT B-scan is used in this method.

Why y-axis is better. In fact, the ablation result, which shows y-axis is better, is not in line with our expectations. The reason is that its 2D version, which takes OCTA projections as input, is equivalent to treating the z-axis as a channel. So we investigated this issue and identified an explanation. In typical convolutional networks, the first convolution layer reduces the spatial resolution by a factor of 2 while significantly increases the number of channels, for instance, from 3 to 64. Consequently, there is no significant loss of information in this layer. In contrast, our additional convolution blocks drastically reduce the number of channels, from 256 to 64, resulting in a loss of information if they are not appropriately trained. When considering the z-axis as channel dimension, this loss of information is especially significant. However, it is less pronounced when using the y-axis as the channel because consecutive B-scans are often similar to each other and contain a lot of redundancy.

Projection supervision. This analysis leads to a method further enhancing the performance, whose key idea is to improve the training of shallow feature extractor. To achieve this, we propose to add another branch onto the EfficientNet backbone, as illustrated in Figure 3.1 (c). This newly added branch functions in a similar way to the decoder of the Unet [55] and is capable of generating OCTA projections from the 3D OCTA volume. By doing so, the additional convolutional blocks, along with some shallow layers in EfficientNet, can better preserve the information necessary for displaying vessel patterns and aiding in AMD grading. It is worth noting that this branch serves only for loss calculation and can be discarded during the inference stage. As a result, we improve accuracy without requiring additional inputs or incurring extra inference time costs.

3.4 Experimental Results

Dataset. Because there is no public OCTA dataset suitable for AMD stage grading, we use our own dataset collected from Jacobs Retina Center at Shiley Eye Institute in experiments.

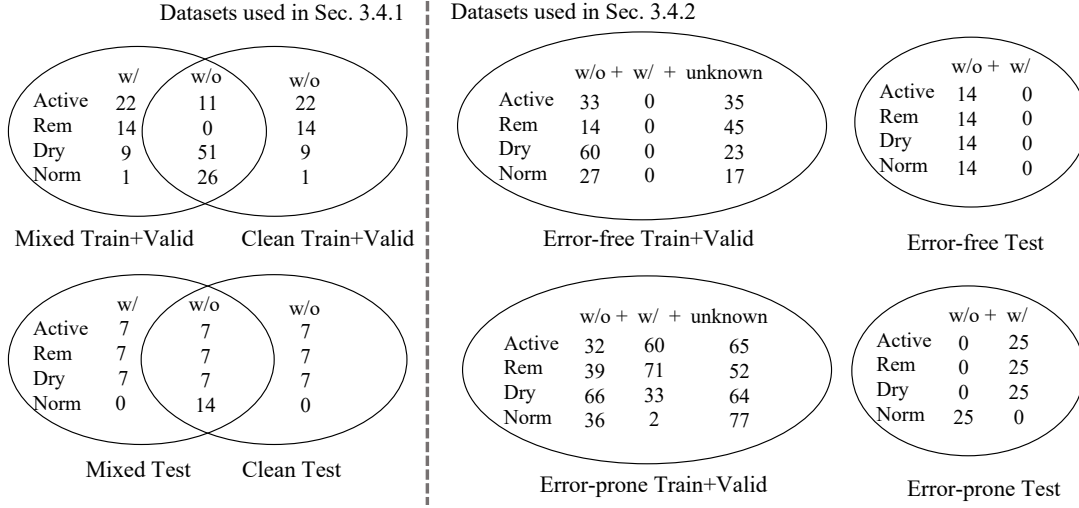


Figure 3.4: Illustration of the datasets used in our experiments

For details of data collection, please refer to Section 2.3.2. The dataset consists of 889 raw OCTA volumes with corresponding projections belonging to four AMD stages: active, remission, dry and normal. Please refer to Figure 1.1 for examples. ‘Active’ means the pathogenic vessels are leaking fluid while ‘remission’ means the pathological vessels were once active but recovered after treatment and showing no fluid. ‘Dry’ represents an early stage of AMD which is not exudative and ‘normal’, as name implies, is obtained by imaging healthy retina. For dataset division, we firstly choose a predetermined number of samples from each category to form the testing set. Then, we randomly select validation set from the remaining samples to conduct a 5-fold validation experiment. Following this strategy, we created two sub-datasets: an easier subset which only had samples with no layer segmentation errors in its testing set, indicated as ‘error-free’ and a harder subset containing numerous samples with errors in its testing set, indicated as ‘error-prone’. Figure 3.4 shows the data amount of datasets used in our experiments. ‘w/’ and ‘w/o’ denote samples with and without layer segmentation errors respectively. ‘Unknown’ refers to samples with no available error annotations. On the left, to keep the data amount same, we created the ‘Mixed’ sub-dataset by replacing error-free samples in the ‘Clean’ sub-dataset with samples containing errors. On the right, the error-free test set only contains samples without

errors, while the error-prone test set includes samples with errors in all AMD stages. For the error-free training set, we removed all known samples with errors to build a relatively error-free set while maintaining the size of training set. For the error-prone training set, all samples except those used for testing were included.

Implementation. We implement all our deep classifiers on PyTorch platform. To save GPU memory, we downsample OCTA projections and volumes to 256×256 and $256 \times 256 \times 256$, respectively. Then we adopt several data augmentations to increase their diversity. In detail, we use random flipping, rotation and cropping with resizing. We randomly apply gamma transformation and Gaussian smooth to increase the diversity of intensity. For projections, we also use grid distortion to augment the shapes. For both 2D and 3D data, we adopt a sample-wise normalization to whiten the sample intensity. Oversampling training data in each category is used to balance their distribution. The networks are trained by Adam optimizer with 10^{-5} weight decay. The initial learning rate is 10^{-3} and decreases via a cosine scheduler with minimum value 10^{-5} . The cosine loss serves as our optimization target, which is proven to be effective with small data amounts [56]. For projection supervision branch, we employ MSE to compute projection differences, and the ratio between cosine loss and MSE is decided by ablation experiments shown in Table 3.5. We have released our models and codes on this website.

3.4.1 Influence of layer segmentation errors

We create two datasets to assess the impact of layer segmentation errors: ‘clean’ and ‘mixed’. They have the same size but the ‘clean’ set only includes error-free samples, while the ‘mixed’ set includes data with and without segmentation errors. For ‘clean’ dataset, we randomly selected 14 samples from each category for testing and used the remaining samples as training. Then we considered samples with errors. For the three categories except ‘normal’, we replaced 7 testing samples with randomly selected 7 samples with errors. Consequently, we obtained a testing set that has the same scale as ‘clean’ but includes data both with and without errors. We

Table 3.2: Classification accuracy with different training/testing datasets. ‘Clean’ means a set with no segmentation errors and ‘Mixed’ means a set mixed with samples with and without errors.

Train on	Test on	Accuracy
Clean set	Clean set	69.64%
Clean set	Mixed set	53.57%
Mixed set	Clean set	64.29%
Mixed set	Mixed set	57.14%

generated a training set with same properties by running the same process and name this dataset as ‘mixed’. Transferring from ‘mixed’ to ‘clean’ dataset, we can simulate the process in which we correct layer segmentation errors in ‘mixed’ dataset.

We conduct 5-fold validation experiments using Resnet18 [57] on both datasets and use the ensemble prediction as the final result by averaging the predictions of 5 classifiers trained in each fold. Note that we can choose to train and test with either ‘clean’ or ‘mixed’ set, resulting in 4 different combinations, shown in Table 3.2. The first two rows of Table 3.2 show that the classifier struggles to generalize from clean samples to those with errors, indicating data with and without errors follow different distributions. Taking the last row into account, we find adding samples with errors to the training set benefits, showing that the classifier may learn the joint distribution of samples with and without errors if given enough training data. The accuracy in the last two rows shows that, even trained on samples with errors, the clean test still works better, implying that samples with errors are hard to learn.

Methodology from experiments. This experiment suggests two ways for improving the performance: 1) collecting enough data to cover the joint distribution of samples with and without errors; 2) avoiding layer segmentation errors and reducing the gap between each distribution. We focus on the second option in this chapter and explore in the other direction in later chapters.

3.4.2 Performance of deep classifier

In this section, we experiment mainly on two datasets, namely error-free and error-prone. For the error-free test set, we utilized the clean test set from Sec. 3.4.1. However, as indicated in Table 3.2, the training set must be cleaned to enhance its performance. Therefore, we integrated error-free training samples along with samples without error annotations, referred to as ‘unknown’ samples in Figure 3.4, while eliminating all known error-prone training samples. By adopting this approach, we can effectively cleanse the training set while keeping its size. In contrast, the error-prone test set comprises solely of samples containing errors in all AMD stages, and all samples except those designated for testing were utilized to construct the error-prone training set.

For baseline method, as far as we know, there is no deep learning based AMD stage grader using OCTA only. Therefore, we use a multimodal AMD grader [51] for performance comparison. We train their networks on our dataset for fair comparison since their dataset is not publicly available. We implement two classifiers based on their official codes which use OCTA information only and use multimodal information from OCT B-scan, OCT and OCTA projection. Note that there are two differences between their task and ours: 1) they do not have ‘remission’ in their target categories, and 2) we do not have high-definition OCT B-scans in our dataset, so we use common B-scans as an alternative. We also replace ORCC projection used in their experiments with SVC projection. We conduct 5-fold validation experiments on two sub-datasets: an easier subset which only has samples with no layer segmentation errors in its testing set (error-free), and a harder subset containing numerous samples with errors (error-prone). In Table 3.3 and 3.4, we report the ensemble accuracy, AUC of RoC in ‘one v.s. rest’ manner, and the performance of human experts on the same test sets. Please refer to supplement material for details of human expert evaluation.

As shown in Table 3.3, Ours-2D with pretrained weights significantly outperforms Thakoor et. al. [51] regardless of the use of multimodal information. This is due to the difference in network structure and training strategy. Note that [51] trained a customized network

Table 3.3: Ensemble accuracy (%) and RoC-AUC performance of different AMD graders with 2D inputs. Error-free and Error-prone are two testing sets w/ and w/o segmentation errors, respectively. MM: Multimodal information (including OCT B-scan, OCT and OCTA projections), PT: Pretraining.

2D Input	Setting		Error-free		Error-prone	
	MM	PT	Accuracy	AUC	Accuracy	AUC
Thakoor	✗	✗	55.36	0.8159	57	0.8176
et. al. [51]	✓	✗	62.5	0.8512	66	0.8428
ours(2D)	✗	✗	73.21	0.8565	62	0.8065
	✗	✓	80.36	0.9264	72	0.8697
Human	-	-	58.92	-	60	-

Table 3.4: Ensemble accuracy (%) and RoC-AUC performance of different AMD graders with 3D inputs. PT: Pretraining, PS: Projection Supervision.

3D Input	Setting		Error-free		Error-prone	
	PT	PS	Accuracy	AUC	Accuracy	AUC
Effic.Net 3D	✗	✗	75	0.9489	69	0.8841
Med.Net34 [58]	✓	✗	73.21	0.9238	73	0.9009
ours(3D)	✓	✗	82.14	0.9524	74	0.9055
	✓	✓	83.93	0.9298	80	0.912

with four 3D convolution and three fully connected layers from scratch, which is much simpler than EfficientNet. The benefit of EfficientNet backbone is evident from the first and third rows and, as shown in the third and fourth rows, pretraining the model further improves its ability to identify useful patterns in OCTA projections. Note that Ours-2D demonstrates significant improvements compared to human experts, indicating the potential of OCTA as a diagnostic modality in AMD grading. These promising results call for further exploration of OCTA-derived biomarkers for accurate AMD diagnosis.

When considering Ours-3D in Table 3.4, we observed a notable improvement compared to Ours-2D. Since the structures of both networks are quite similar (Figure 3.1), this gain demonstrates the advantages of directly grading 3D OCTA volumes and reducing the gap between data with and without errors. The advantage of Ours-3D method can be also substantiated by examining the performance differences of Ours-2D and Ours-3D in error-free and error-prone

settings. In the error-free setting, where fewer samples are affected by errors, the improvement gained from using Ours-3D is relatively smaller. However, in the error-prone setting, where errors are more prevalent, the performance of Ours-2D experiences a significant decline, while Ours-3D maintains high performance levels. This differential behavior in error-free and error-prone settings serves as evidence that the proposed Ours-3D method is more robust in the presence of layer segmentation errors.

In comparing Ours-3D with 3D EfficientNet and MedicalNet34 [58], both of which utilize 3D convolutions, we find that 2D backbone is more effective. This finding is actually consistent with some early works in action recognition [59, 60, 61]. Their experiments verified that well-designed 2D convolution network is better than 3D, especially when training data is limited. Our result indicates that utilizing a pretrained 2D network is currently a promising method for analyzing 3D OCTA until a large-scale 3D OCTA dataset is available. Finally, the efficacy of our proposed projection supervision is demonstrated in the last two rows, where the accuracy is improved to over 80%. It also indicates that OCTA is an informative modality for AMD grading.

3.4.3 Ablation study

This section presents our ablation experiments, which aim to investigate the impact of different factors on the performance of our classifiers. Specifically, we examine the effects of different choices of channel axis, different ratios of loss weights, and the use of pretrained weights and projection supervision. The accuracy of different classifiers trained on the first validation fold are reported in Table 3.5.

Firstly, our results indicate that taking y-axis as the channel is more effective than z-axis when projection supervision is not used. The reason has been elaborated in Sec. 3.3.3. Then the use of projection supervision improves the z-axis inputs while negatively impacting y-axis channel inputs. This outcome is consistent with our expectations since taking z-axis as channel means taking x and y dimension as spatial which aligns with the spatial dimension of OCTA projections.

Table 3.5: Ablation experiments w.r.t the choice of channel axis and the loss weight ratio. The accuracy here pertains to the performance of individual classifier trained on the first validation fold, instead of the outcome of ensemble.

Settings				Accuracy (%)
Channel axis	Pretrain	Proj. Supervision	Weight ratio	
y axis				54
y axis	✓			69
y axis	✓	✓	1:10	64
z axis				50
z axis	✓			64
z axis	✓	✓	$1:10^{-1}$	69
z axis	✓	✓	1:10	72
z axis	✓	✓	$1:10^3$	74
z axis	✓	✓	$1:10^4$	67

It is unreasonable to generate OCTA projections from a stack of OCTA B-scans. Furthermore, our experiments on various weight ratios demonstrate that the ideal ratio between Cosine loss and MSE loss is approximately $1:10^3$ for z-axis channel inputs. Finally, our experiments also show that the use of pretrained weights improves the performance of the classifiers, regardless of which dimension is selected as the channel.

3.4.4 Detailed comparison with human expert

As described in Figure 1.1, we expect our method outperforms human experts in this four-stage grading task. To compare the performance of our proposed method with that of human experts, in this section, we conducted human expert evaluation and provided a detailed analysis of the confusion matrix in different settings.

Human expert evaluation

Evaluation of OCTA images was performed by two masked retina specialists who regularly employ OCTA in clinical settings. These specialists assessed the presence or absence of CNV on OCTA en-face projections only and remained blind to other images (SD-OCT B-scan, OCT

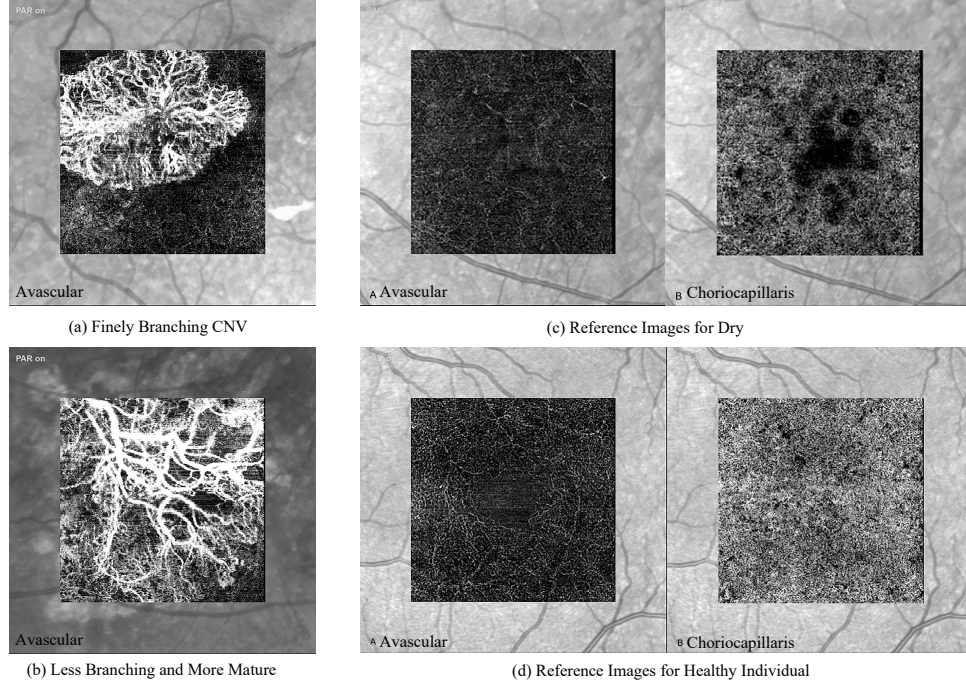


Figure 3.5: Standardized reference images used in human expert evaluation experiment.

en-face scan), as well as clinical information. It's crucial to note that these graders were different from the retina specialists engaged in data labeling.

To predict the clinical category (Active, Remission, Dry AMD, and Normal), the masked graders focused exclusively on OCTA en-face projections, including the superficial vascular complex, deep vascular complex, avascular layer, and choriocapillaris. Initially, the presence or absence of CNV was determined using standardized reference images, comprising en-face structural OCTA images of the superficial inner retinal plexus, deep inner retinal plexus, outer retina, and choriocapillaris for each eye. Images indicating an evident neovascular plexus in the avascular layer and/or choriocapillaris were interpreted as indicating CNV presence. Subsequently, if vessels were detected, graders scrutinized their morphology, attempting to classify them as finely branching (presumably active, as depicted in Fig. 3.5a) or less branching and more mature (predominantly inactive, as illustrated in Fig. 3.5b). This classification was based on characteristics previously described by Coscas et al [9]. for reference, aiding in distinguishing

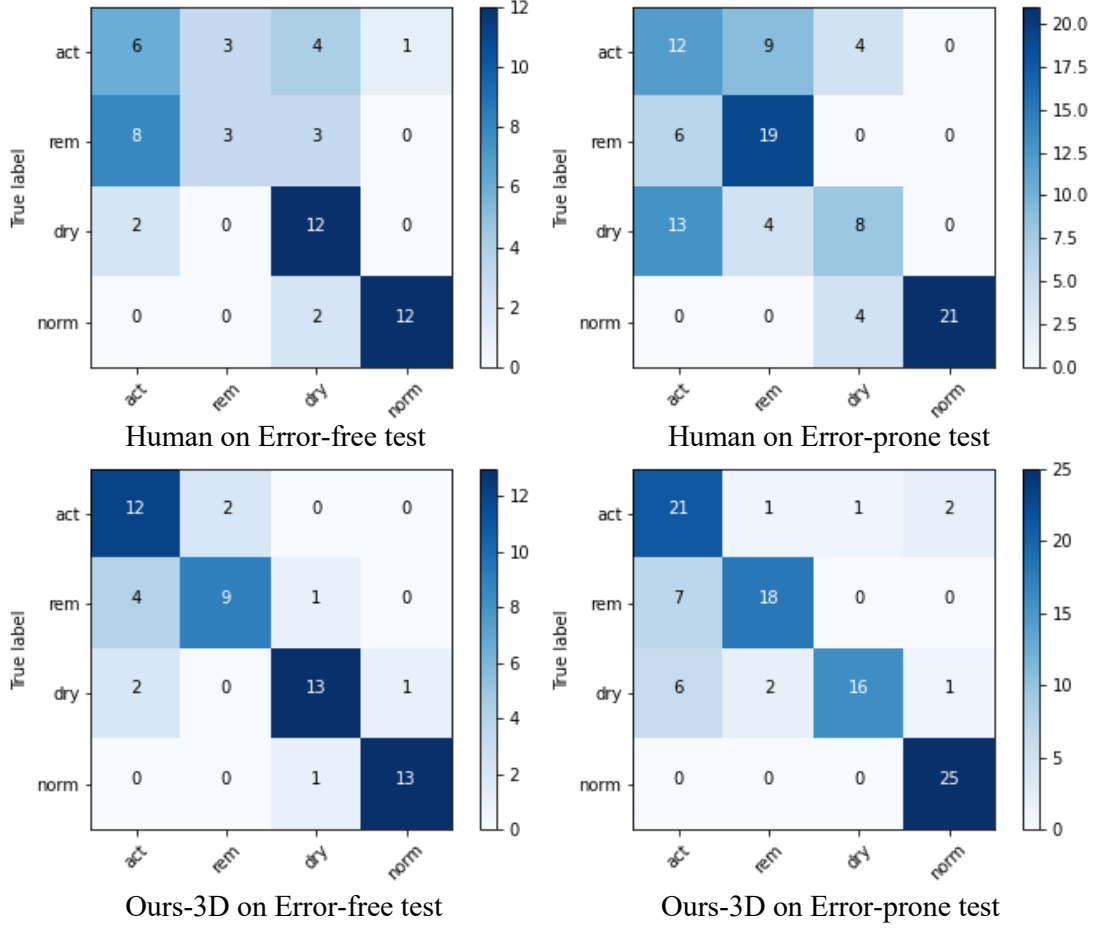


Figure 3.6: Confusion matrix comparison between our proposed method and human experts on different test sets. Ours-3D, outperforms human experts in accurately distinguishing between the ‘active’ and ‘remission’ categories. Also, as indicated by the smaller performance drop observed in the ‘dry’ category, our method demonstrates greater robustness to layer segmentation errors.

the clinical category. Additionally, reference images for other categories were employed, such as Fig. 3.5c, showcasing OCTA in a patient with dry AMD, as characterized by Arya et al.[62]. Fig. 3.5d demonstrated a healthy individual with normal choriocapillaris and avascular layer in the OCTA en-face projection. These references, coupled with their clinical acumen, enabled the two expert graders to arrive at a consensus grading for predicting the four categories. Remarkably, the same 56 (or 100) samples that the network classifiers used for testing were utilized by these two masked human graders to predict the category.

Result analysis

Figure 3.6 reported the confusion matrices of our proposed classifiers and human experts on different test sets, respectively. Firstly, we evaluated the matrix of the human expert on the error-free test set. It can be observed that the ophthalmologist who took this experiment performed well in distinguishing between the ‘dry’ and ‘normal’ categories but struggled in differentiating between the ‘remission’ and ‘active’ categories. This highlights the ongoing challenge in accurately determining the active stage of AMD for human experts, thereby emphasizing the significance of our work.

Subsequently, we analyzed the matrix of the human expert on an error-prone test set. It can be found that the human expert continued to face difficulty in distinguishing between the ‘active’ and ‘remission’ categories, but this time, the accuracy of the ‘dry’ category significantly decreased. This is exactly the consequence caused by layer segmentation errors, i.e. the incorrect vascular networks and missing vessels in the OCTA projections caused confusion for the human expert.

In contrast, our proposed method, termed Ours-3D, shows a significant improvement in the confusion matrix, accurately classifying the majority of test samples in each category. On the error-free test set, Ours-3D performed slightly worse in the ‘remission’ category, owing to the relatively fewer training samples in this category. On the error-prone test set, our method demonstrated greater robustness to segmentation errors by directly taking the raw OCTA volume as input and bypassing the impact of those errors. Overall, our proposed method not only outperforms human experts in this AMD grading task but also offers increased robustness to segmentation errors, which is a critical consideration in accurately detecting and grading AMD.

3.5 Discussion

In this chapter, we firstly elaborate the influence of layer segmentation errors in the context of AMD stage grading and propose to address it via analyzing the 3D OCTA volume directly. With the pretrained 2D EfficientNet backbone and projection supervision, we achieve an accuracy of over 80% on both error-free and -prone test sets, which significantly outperforms 60% accuracy of human experts. Our results suggest that OCTA modality alone can identify different AMD stages and encourage the exploration of OCTA-derived biomarkers for diagnosis.

Although achieving the best grading performance, 3D volume input has its own limitations. For one example, it needs more computation resources and the I/O is much slower. For another example, not every OCTA instrument supports raw OCTA volume exportation, hindering the broader application of a 3D classifier. So in following chapters, we explore in the other direction: expanding the training dataset.

Chapter 3, in part, is a reprint of the material as it appears in 2023 International Conference on Computer Vision (ICCV) Workshop [3], Haochen Zhang, Anna Heinke, Carlo Miguel B. Galang, Daniel N. Deussen, Bo Wen, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Cheolhong An. “Robust AMD Stage Grading with Exclusively OCTA Modality Leveraging 3D Volume”, in *ICCV* workshop, 2023. The dissertation author is the primary author of this paper. It, in part, is also a reprint of the material as it appears in *RETINA* [4], Anna Heinke, Haochen Zhang, Daniel N. Deussen, Carlo Miguel B. Galang, Alexandra Warter, Fritz G. P. Kalaw, Dirk-Uwe G. Bartsch, Lingyun Cheng, Cheolhong An, Truong Q. Nguyen, and William R. Freeman. “Artificial Intelligence for OCTA-based Disease Activity Prediction in Age-Related Macular Degeneration”, *RETINA*, 2022. The dissertation author is a co-primary contributor to this work responsible for AI classifier development and experiment results analysis.

Chapter 4

Cross-instrument OCTA Translation with Class Condition

According to the experiments shown in Table 3.2 in the previous chapter, we present two strategies to handle the problem of layer segmentation errors: one involves direct analysis of the 3D raw OCTA volume, thereby mitigating the influence of segmentation errors, while the other entails expanding the training dataset to encompass the data distribution affected by segmentation errors. Given the impracticality of exporting 3D raw data for OCTA instruments like Optovue and Topcon, hindering the broader application of a 3D classifier, this chapter explores the latter approach.

4.1 Introduction

The retina, a vital part of the human eye, transforms light into neural signals through a photosensitive layer of optic nerve tissue. At the core of retina lies the macula, essential for detecting color and light intensity, without which detailed and vivid color perception is compromised [7]. Unfortunately, various eye diseases, such as age-related macular degeneration

(AMD), choroidal neovascularization (CNV), and diabetic macular edema (DME), can impact the health of the macula. AMD, in particular, is a prevalent cause of vision loss, contributing to a significant health concern [8]. Since the progression of AMD often associates with abnormal vascular changes, diagnostic tools detecting pathological vessels is crucial for optimal treatment and vision preservation.

Optical coherence tomography angiography (OCTA), an emerging imaging technique, facilitates visualization of retinal and inner choroidal circulation without dye injection, becoming crucial for detecting CNV lesions in neovascular AMD [25]. OCTA offers quantitative metrics like foveal avascular zone area and capillary network density, aiding in the assessment of retinal diseases and morphological changes in vascular networks responding to treatments [20]. Although not yet a gold standard in AMD treatment decision-making, the non-invasive nature of OCTA and its ability to detect subclinical changes position it as an ideal diagnostic and monitoring tool.

Recent research [63] highlighted the limited capability of en-face OCTA images in identifying CNV, and OCTA alone struggled with accurate CNV activity determination. To unlock the full potential of OCTA in clinical applications, more effective analysis methods are needed. Given the superior performance of deep neural networks (DNNs) in surpassing human abilities in natural image classification [12], our study focuses on developing deep classifiers for AMD stage grading *based solely on OCTA images*. These classifiers directly learn informative features from the data, proving more efficient than humans in discerning correlations within multiple projections, en-face OCTA images at different depths [4].

To enhance the performance of a deep classifier in AMD stage grading, a recent investigation [3] scrutinized en-face OCTA images produced by the Heidelberg OCTA instrument, identifying a substantial impact from segmentation errors of retina layers. In [3], two strategies were proposed to handle these errors: one involves direct analysis of the 3D raw OCTA volume, thereby mitigating the influence of segmentation errors, while the other entails expanding the training dataset to encompass the data distribution affected by segmentation errors. Given the im-

practicality of exporting 3D raw data for OCTA instruments like Optovue and Topcon, hindering the broader application of a 3D classifier, this chapter explores the latter approach. To facilitate this exploration, we amassed an additional OCTA dataset generated by the Optovue OCTA instrument and trained deep classifiers using cross-instrument data, leveraging style transfer techniques.

In detail, as the two datasets comprise distinct patient groups, we employ the CycleGAN framework [15] for unpaired style transfer. Utilizing a well-trained CycleGAN translator facilitates the conversion of Heidelberg samples to Optovue and vice versa, introducing additional contents into each domain. This process also enhances sample diversity within each domain through cycle consistency and identity outputs. Unlike many existing style transfer methods that focused on realistic visual quality, our primary objective is enhancing classification accuracy. To optimize machine recognition quality of transferred images, we introduce additional constraints using a pretrained deep classifier. This constraint could be implemented in either fully-supervised or unsupervised manners: in domains with available labels, we ensure the transferred image belongs to the same class; in the absence of labels, we aim to make the representatives of the image before and after transferring as similar as possible. Experimental results demonstrate that the CycleGAN translator effectively enhances classification performance in each domain, and our proposed class-conditioned approach further boosts the accuracy.

4.2 Related Work

4.2.1 OCTA analysis

As discussed in Chapter 3.2.1, we have reviewed deep learning-based OCTA analysis. In this section, we highlight a few significant existing works to illustrate the distinctions of our approach. As OCTA enables the visualization of the retinal vascular network, the majority of OCTA-based studies have concentrated on segmentation tasks [32, 34]. For example, Alam et

al. [31] utilized U-Net for artery-vein classification, employing transfer learning to address limited dataset challenge. Besides segmentation, some OCTA studies delve into deep learning-based classification, such as Le et al. [35] which used the VGG16 network for diabetic retinopathy stage grading. Lin et al. [37] incorporated boundary shape and distance maps for simultaneous classification and segmentation. To the best of our knowledge, only one previous study focused on AMD diagnosis using OCTA exclusively. In Zhang et al.’s work [3], they investigated the impact of layer segmentation errors and proposed a 3D OCTA volume network with a 2D EfficientNet backbone, trained with additional en-face OCTA supervision. Diverging from this approach, our method takes a unique route by collecting a new dataset using a different instrument to broaden training data diversity and, consequently, enhance accuracy.

4.2.2 Neural style transfer technique

Style transfer manipulates the visual appearance of an image by altering its thematic style while preserving its content [64]. A pioneering work by Gatys et al. [65] introduced a seminal neural style transfer algorithm that utilized DNNs to separate and recombine content and style representations. Since then, neural style transfer has become a popular topic in both academic and industrial fields, with numerous works published to improve speed and scalability. Generally, existing methods can be divided into online image optimization methods and offline model optimization methods.

The direction of online image optimization includes works by Gatys et al. [65], and its improved version [66], which accessed image style factors of space, color, and scale during stylization to substantially improve flexibility and quality. Li et al. [67] treated neural style transfer as a domain adaptation problem, theoretically proving the significance of matching the Gram matrices of feature maps. In the same year, Li et al. [68] proposed a Laplacian loss, based on the Laplacian matrix produced by a Laplacian operator, to produce more appealing stylized images with fewer artifacts. The main limitation of this direction is efficiency. In contrast offline model

optimization methods have advantages in speed and computational cost. For example, Ulyanov et al. [69] moved the computational burden to training stage by using a compact feed-forward convolutional network to generate samples of the same texture. In the same year, Johnson et al. [70] proposed an image transformation network with perceptual loss that achieved real-time style transfer. Li et al. [71] proposed Markovian Generative Adversarial Networks for efficient texture synthesis, addressing the efficiency issue. Later, Ulyanov et al. [72] found that instance normalization significantly improved stylization quality compared to batch normalization.

One notable progress is CycleGAN [15], which extended style transfer to unpaired image domains, allowing the adaptation of artistic styles between diverse datasets, for example in area of medical imaging. Although in recent years, diffusion model [73, 74] has shown powerful ability in image generation. It can hardly replace CycleGAN currently because of its specially ability to trained on unpaired dataset. This is also why we use CycleGAN instead of popular diffusion model as translator framework for cross-instrument OCTA translation in Chapter 4.

4.3 Proposed Methods

4.3.1 Overview of CycleGAN framework

Existing image translation methods often depend on paired datasets, where corresponding images in two domains share the same content. However, in our case, the Heidelberg and Optovue datasets contain retinas from different patients, making paired data unavailable. To overcome this limitation, we employ the CycleGAN framework [15] which addresses the absence of paired data by introducing a cycle-consistency constraint. For instance, in the Green cycle of Figure 4.1, a real Optovue sample x undergoes translation to the Heidelberg domain using the generator $G_{O2H}(\cdot)$, resulting in $\hat{y}$. Subsequently, $\hat{y}$ is translated back to the Optovue domain through another generator $G_{H2O}(\cdot)$, producing the recovered Optovue image $\tilde{x}$. By imposing an L1 loss between x and $\tilde{x}$, we ensures faithful reconstruction, preserving both content and style in Optovue domain.

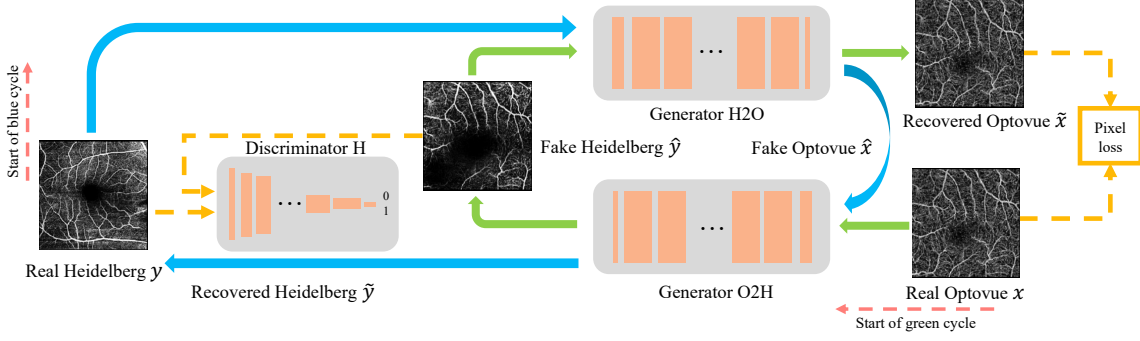


Figure 4.1: Illustration of our framework utilizing CycleGAN to transfer samples between Heidelberg and Optovue domains. For simplicity, only one channel of OCTA projection is displayed, and the identity loss is omitted. For more details please refer to Figure 4.2

Adversarial loss between y and $\hat{y}$ enforces the transferred image share same style as samples in Heidelberg domain. An analogous procedure is employed for the blue cycle, and both cycles share the generators $G_{O2H}(\cdot)$ and $G_{H2O}(\cdot)$. Thus in this manner, a robust and accurate mapping can be learned even without paired training samples.

Additionally, following the recommendation in [15], we incorporate an identity loss, which as depicted in Figure 4.2, measures the L1 difference between x and $G_{H2O}(x)$, along with the L1 difference between y and $G_{O2H}(y)$. It plays a pivotal role in preserving channel relationships, which is critical in our en-face OCTA data. In summary, the vanilla training loss of CycleGAN without class constraint is:

$$\begin{aligned}
 l_{cyclegan}(x, y) = & l_{cyc}(x, \tilde{x}) + l_{cyc}(y, \tilde{y}) \\
 & + \alpha(l_{GAN}(x, \hat{x}) + l_{GAN}(y, \hat{y})) \\
 & + \beta(l_{idt}(x, G_{H2O}(x)) + l_{idt}(y, G_{O2H}(y)))
 \end{aligned} \tag{4.1}$$

where α, β are loss weights.

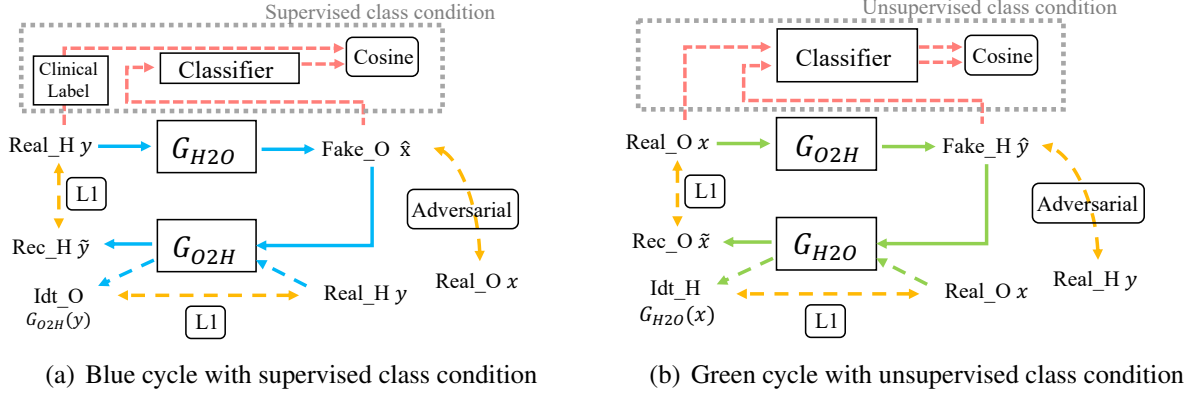


Figure 4.2: The disassembly diagram of Figure 4.1 showing the blue (a) and green (b) cycles, respectively, with the proposed supervised (a) and unsupervised (b) class constraints.

4.3.2 Explicit class condition

Diverging from most existing style transfer works, our primary objective is facilitating the classifier training with transferred images, rather than emphasizing realistic visual quality. Prior research [75] has established that, in image restoration area, optimal classification performance is not achievable when images undergo restoration with distortions or perceptual losses alone. We extend this principle to style transferring. Therefore in this section, we investigate explicit constraint to elevate the machine recognition quality of transferred images. Leveraging a pretrained classifier, we propose implementing this constraint through either a fully-supervised approach, as demonstrated in Figure 4.2 (a), or an unsupervised approach, as illustrated in Figure 4.2 (b), details of which will be presented in the following paragraphs.

Supervised class constraint

In the ideal scenario where class annotations are available, a straightforward approach to incorporate class constraints involves applying a pretrained classifier $cls(\cdot)$ to the transferred image and enforcing it to remain within the same class. For example, illustrated in Figure 4.2 (a), we feed the fake Optovue image $\hat{x}$ into the classifier and then compute the classification loss with respect to the label of real Heidelberg image y . We use cosine loss rather than cross

entropy, as it has demonstrated effective with limited data [56]. It is worth noticing that the classifier is pretrained and frozen during CycleGAN training process. Thus, to minimize the additional classification loss, the generator $G_{H2O}(\cdot)$ must learn to preserve category information while performing style transfer. Combined with the vanilla loss, the final CycleGAN training loss with supervised class constraints is:

$$l_{total} = l_{cyclegan}(x, y) + \gamma(l_{cos}(label(x), cls(\hat{y})) + l_{cos}(label(y), cls(\hat{x}))) \quad (4.2)$$

where γ is loss weight, $cls(\cdot)$ is a pretrained classifier.

Unsupervised class constraint

Alternatively, this class constraint can be implemented in an unsupervised manner when labels are unavailable in certain cases. As illustrated in Figure 4.2 (b), we apply the same classifier to the real Optovue image x to generate a pseudo label for the classification loss. This pseudo label, akin to the soft labels utilized in teacher-student distillation learning approaches [76], encapsulates essential category information. Moreover, this approach allows for computing classification loss at the feature level, wherein the classifier is replaced with a feature extractor. In the scenario presented in Figure 4.2 (b), the classifier extracts the image representations for both real Optovue image x and the transferred fake Heidelberg image $\hat{y}$. Despite differences in style, these images share identical semantic content. Consequently, we utilize a cosine distance as loss function to make their image representatives as close as possible. When integrated with the vanilla loss, the final CycleGAN training loss incorporating unsupervised class constraints is:

$$l_{total} = l_{cyclegan}(x, y) + \gamma(l_{cos}(cls(x), cls(\hat{y})) + l_{cos}(cls(y), cls(\hat{x}))) \quad (4.3)$$

where γ is loss weight, $cls(\cdot)$ is a pretrained feature extractor.

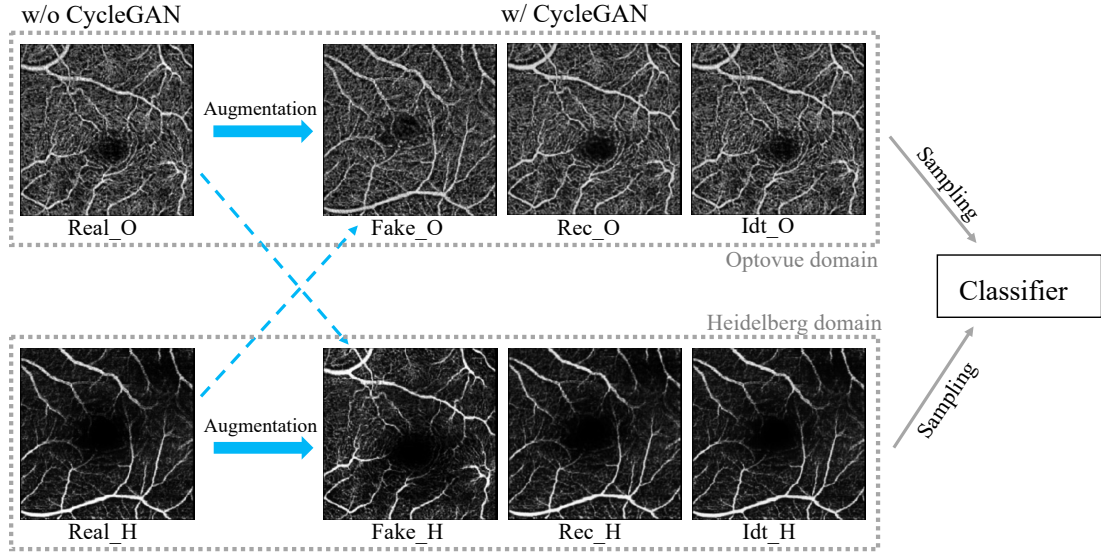


Figure 4.3: Illustration of the CycleGAN-based data augmentation for DNN classifier training. Synthetic samples are generated via CycleGAN as shown in Figure 4.2. The classifier can be trained with real or synthetic data from each domain, employing predetermined sampling strategies.

4.3.3 Classifier training with CycleGAN

In this section, we outline the process of training a classifier using the augmented training set generated by CycleGAN. Leveraging a well-trained CycleGAN translator, trained with the proposed class condition, we acquire the capability to transfer Heidelberg samples into Optovue style and vice versa. This augmentation introduces samples with previously unseen content in each domain, illustrated by ‘Fake_O’ and ‘Fake_H’ in Figure 4.3. Despite sharing the same style as real samples, these synthetic data exhibit entirely different vascular patterns, enriching the diversity of the training dataset.

Furthermore, the CycleGAN translator allows for intra-domain sample augmentation by utilizing the outputs of cycle consistency and identity, exemplified by ‘Rec_O’ and ‘Idt_O’ in the Optovue domain. While in an ideal scenario ‘Real_O’, ‘Rec_O’, and ‘Idt_O’ would be identical (if the loss $l_{cyclegan}$ reaches zero, which is practically unattainable), they remain slightly different but adhere to the distribution of the Optovue domain. Analogous to conventional data augmentation

techniques like random distortion, these variations hold the potential to enhance classifier training, particularly when combined with our proposed class constraint. Utilizing different predetermined sampling strategy, the classifier can be trained with real or synthetic data from the Heidelberg domain, the Optovue domain, or a combination of both.

4.4 Experimental Results

4.4.1 Implement details

Dataset

In addition to the Heidelberg OCTA dataset utilized in Section 3.4.2, we collected an extra Optovue Solix OCTA dataset from Medical University Gdańsk, Poland. This dataset comprises of 580 en-face OCTA projections, sharing the same target categories—active, remission, dry, and normal—as the Heidelberg dataset. Please refer to Section 2.3.2 for clinical definitions of each category. For dataset division, we initially select 15 samples per category for Optovue and 25 samples per category for Heidelberg to form the testing set. Then, for classifier training, we randomly choose a validation set from the remaining samples to conduct a 5-fold validation experiment. In CycleGAN training, since it is an unpaired translator, we use both the training and validation sets. However, no testing samples are involved in the training process for either network.

Implementation

We implement all DNNs using the PyTorch platform. For CycleGAN, we compute adversarial loss using LSGAN [77] in a PatchGAN manner, with the loss weights α and β consistent with the original paper [15]. The class constraint weight γ is set to 0.1. The networks are trained using the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. All generators and discriminators

undergo 200 epochs with a learning rate of 0.0002, followed by a linear decrease over the next 600 epochs. As for the classifier, we maintain the training settings exactly the same as in [3]. This includes the use of Adam optimizer with 10^{-5} weight decay, an initial learning rate of 10^{-3} decreasing via a cosine scheduler with a minimum value of 10^{-5} . The loss function is the cosine loss [56].

4.4.2 Classification performance

Given the limited instances of OCTA based AMD stage graders, especially on the Optovue OCTA dataset, we implement the recently published 3D Heidelberg classifier [3] and apply it to both Heidelberg and Optovue en-face images. We also report the 3D performance on Heidelberg as a reference. Additionally, we train a multimodal AMD grader [51] as well, but restrict it to use the OCTA modality only due to the absence of OCT B-scans in our Optovue dataset.

Based on the different sampling strategies illustrated in Figure 4.3, we have various experiment settings to consider:

1. Train the classifier with either ‘Real_H’ or ‘Real_O’ exclusively or a mixture of both without CycleGAN augmentation;
2. Train the classifier with a combination of both domains, augmented by CycleGAN, encompassing all eight versions depicted in Figure 4.3;
3. Train the classifier within one instrument domain, augmented by CycleGAN, for example using ‘Real_O’, ‘Fake_O’, ‘Rec_O’ and ‘Idt_O’ for Optovue domain.

By combining all these cases with the type of class condition employed during CycleGAN training, we present the results in Table 4.1 and Table 4.2, where the accuracy of each classifier is evaluated on Heidelberg and Optovue test sets, respectively.

Let’s firstly examine the classifier without CycleGAN augmentation in Table 4.1. Trained within a single domain, the classifier demonstrates optimal performance on the corresponding

Table 4.1: Classification Accuracy without CycleGAN augmentation. Real- $\{H,O\}$ means a mixture of Real_H and Real_O. Real_H 3D, which uses 3D data as input, is provided for reference.

Backbone	Training sample	Target domain	Test Acc (Heid/Opto)
CustomCNN [51]	Real_O	Opto only	0.31 / 0.45
	Real_H	Heid only	0.56 / 0.3
EfficientNet [3]	Real_O	Opto only	0.33 / 0.7333
	Real_H	Heid only	0.72 / 0.5167
	Real- $\{H,O\}$	Heid+Opto	0.67 / 0.7
	Real_H 3D	Heid only	0.8 / -

Table 4.2: Classification Accuracy with CycleGAN augmentation. Please refer to Figure 4.3 for example of different training sample type.

Training sample	Target domain	Class condition	Test Acc (Heid/Opto)
Real_O, Rec_O Fake_O, Idt_O	Opto only	No	0.51 / 0.7167
	Opto only	Supervised	0.55 / 0.8
	Opto only	Unsupervised	0.64 / 0.7833
Real_H, Rec_H Fake_H, Idt_H	Heid only	No	0.73 / 0.65
	Heid only	Supervised	0.78 / 0.6
	Heid only	Unsupervised	0.77 / 0.6833
Real- $\{H,O\}$, Rec- $\{H,O\}$ Fake- $\{H,O\}$, Idt- $\{H,O\}$	Heid+Opto	No	0.68 / 0.7333
	Heid+Opto	Supervised	0.76 / 0.7667
	Heid+Opto	Unsupervised	0.74 / 0.75

test set but significantly underperforms on the other. Simply combining the two domains helps mitigate imbalanced performance but achieves suboptimal outcomes on both test sets. In Table 4.2, introducing CycleGAN augmentation to the two-domain mixture training improves performance on both test sets, particularly with our proposed class constraint. As observed using CycleGAN trained with the supervised class constraint, the classifier achieves optimal performance on both test sets, even slightly surpassing domain-specific training without CycleGAN. When applied to domain-specific training, the CycleGAN translator further enhances accuracy. In Heidelberg domain with CycleGAN augmented data, the performance significantly improves on the Heidelberg test set, approaching the performance with 3D raw data. It also improves

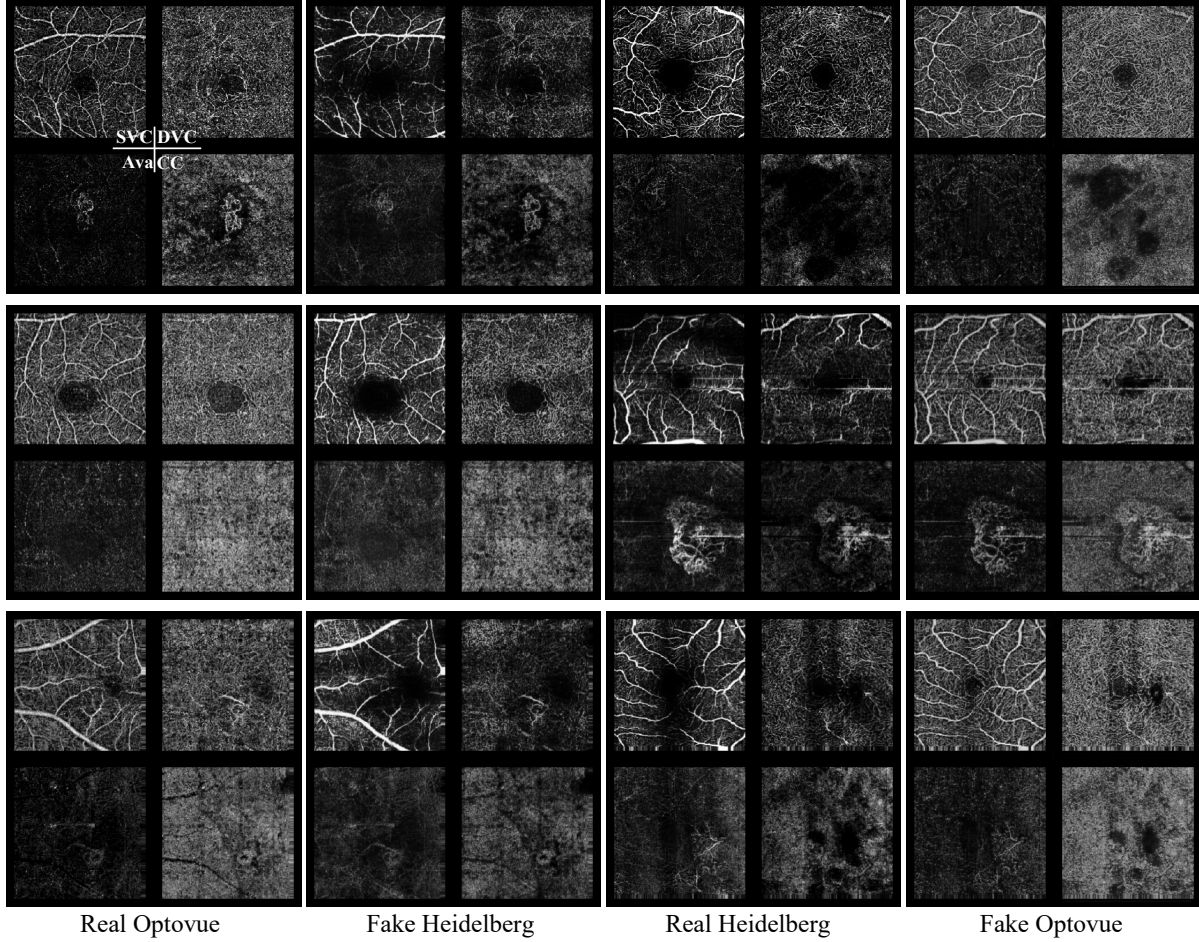


Figure 4.4: Visual inspection of CycleGAN results. Each row shows one example. As can be observed, CycleGAN successfully converts the image style while preserving the vascular patterns.

performance on the Optovue test set, although not optimally. A similar case can be observed in Optovue domain.

These experiments demonstrate the effectiveness of CycleGAN translator as an advanced data augmentation technique in enhancing classification performance. It is noteworthy that in all experiments involving CycleGAN, the performance improvement is marginal, or even worse, without our proposed class constraint. The results empirically affirm that pixel-wise loss is insufficient, and class-related conditions are essential when employing image processing techniques, such as style transferring, to enhance classification performance. Regarding the two

types of proposed class constraints, the supervised one consistently outperforms its unsupervised counterpart, potentially due to the additional label information provided.

4.4.3 Visual inspections

Figure 4.4 presents visual results of the original and transferred OCTA samples for qualitative examination. The real Optovue sample shares vessel patterns with the fake Heidelberg sample, albeit in a different style, and the same is true for the real Heidelberg and fake Optovue samples. In the meantime, real and fake Heidelberg samples share the same style, indicating they belong to the same image domain, as do the real and fake Optovue samples. This results indicate that CycleGAN effectively introduces unseen samples to each domain, augmenting the training set and thereby improving classifier training. Additionally, as seen in the middle row sample, if the real Heidelberg sample has artifacts caused by layer segmentation errors, the transferred fake Optovue sample replicates these artifacts. This feature enriches the Optovue image domain with layer segmentation errors, enhancing the robustness of the Optovue domain classifiers.

4.5 Additional Experiments

In this section, we present additional experiments conducted on the Optovue dataset, focusing on two key aspects: the impact of field of view and the exploration of classifier backbones. While these experiments are not directly tied to the CycleGAN technique, they offer valuable insights into OCTA-based AMD diagnosis, complementing our primary investigation.

4.5.1 Problem of difference in field of view

In Section 2.3.2 as illustrated in Figure 2.4, we discussed the variation in the field of view (FoV) of OCTA projections between our Heidelberg and Optovue subsets, with dimensions of 3x3 mm and 6x6 mm, respectively. Throughout this chapter, we maintained FoV alignment by

Table 4.3: Classification Accuracy without FoV alignment, which means Optovue samples has doubled FoV than Heidelberg samples but share same resolution.

Backbone	Training sample	Target domain	Test Acc (Heid/Opto)
EfficientNet [3]	Real_O	Opto only	0.34 / 0.85
	Real_H	Heid only	0.73 / 0.6
	Real_{H,O}	Heid+Opto	0.78 / 0.7667

centrally cropping Optovue samples and downsampling Heidelberg samples by default. However, in this section, we present the results without FoV alignment in Table 4.3. It’s important to note that a direct comparison between the values in Table 4.1 and those in Table 4.3 is not entirely fair, as we expanded the training set for both the Heidelberg (from 899 to 1478 samples) and Optovue (from 580 to 1003 samples) subsets.

From Table 4.3, two noteworthy observations emerge. Firstly, a striking asymmetrical effect is apparent when integrating data from two distinct instruments: while the accuracy within the Heidelberg domain increases from 73% to 78%, the accuracy within the Optovue domain decreases from 85% to 76.67%. This disparity can be attributed to two primary factors. 1) As elucidated in Section 3.4.1, the Heidelberg dataset exhibits category-dependent segmentation errors, significantly impacting classifier learning. Thus, incorporating Heidelberg data into the training of Optovue data complicates the problem, resulting in suboptimal performance. 2) The two OCTA datasets feature different FoVs, with Heidelberg Spectralis providing 3 x 3 mm images and Optovue scans offering a larger 6 x 6 mm image size. Augmenting Heidelberg data with Optovue data reinforces predictions, as the additional area will not do harm to decision-making. Conversely, augmenting Optovue with Heidelberg data presents challenges, as the missing areas in Heidelberg compel the classifier to overlook valuable cues exclusive to Optovue, complicating the task.

The second observation is that in the Optovue test set, despite being trained solely on Heidelberg data, the classifier achieved a 60% accuracy, twice as effective as the reverse scenario, which attained 34% accuracy. The explanation could also be two key factors. 1) The Heidelberg

Table 4.4: Neural network backbone searching result. EfficientNet-b5 is still the optimal choice considering training speed and computational resource constraints.

Backbone family	Version	Params	Acc (Heid/Opto)
EfficientNet [52]	b4	19M	0.4 / 0.8
	b5	30M	0.34 / 0.85
	b6	43M	0.38 / 0.7833
	b7	66M	0.37 / 0.85
ConvNeXt [78]	base	89M	0.33 / 0.75
	large	198M	0.36 / 0.7667
	xxlarge	350M	0.43 / 0.8
ViT [79]	base	87M	0.35 / 0.7
	large	305M	0.39 / 0.6667

dataset offers a larger training sample size (1478 vs. 1003), mitigating the risk of overfitting and fostering generalization. 2) As discussed earlier, Optovue data encompass a larger FoV, resulting in a performance drop when training on a larger FoV (Optovue domain) and inferring on a smaller one (Heidelberg domain) due to missing spatial clues. In summary, Table 4.3 underscores the impact of FoV on two critical aspects of classifier performance.

4.5.2 Backbone searching for Optovue data

Considering the distinct characteristics of Optovue data compared to Heidelberg, to optimize the classification accuracy of our model, we conducted a network backbone search focusing on three families: EfficientNet, ConvNeXt, and Vision Transformer. EfficientNet [52], renowned for its efficiency in computational resources and model size, achieves state-of-the-art performance across diverse image recognition tasks. ConvNeXt [78] enhances traditional convolutional neural networks by incorporating cross-stage connections, facilitating improved information flow between convolutional layers. Vision Transformers [79] extend transformer architectures from natural language processing to image tasks by dividing images into patches and processing them through transformer blocks, enabling effective capture of long-range dependencies. The results in Table 4.4 underscore the significance of data volume, as larger networks like ConvNeXt and

ViT require more data for effective training. Among EfficientNet variants, b5 and b7 exhibit superior performance, with EfficientNet-b5 being the optimal choice considering training speed and computational resource constraints.

4.6 Discussion

In this study, we employed an unpaired style transfer algorithm, CycleGAN, for cross-instrument data augmentation to enhance the accuracy of AMD stage grading using en-face OCTA images exclusively. The proposed approach not only enhances diversity within each domain but also generates unseen data based on samples from the other domain. Our experiments prove the CycleGAN translator is particularly beneficial if trained with the proposed class-related constraints. We hope this study draws more attention to the significance of semantic-related loss design in image processing algorithms aimed at improving classification as downstream task.

Chapter 4, in part, is a reprint of the material as it appears in the Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBC) [5], Haochen Zhang, Anna Heinke, Krzysztof Broniarek, Carlo Miguel B. Galang, Daniel N. Deussen, Ines D. Nagel, Katarzyna Michalska-Małecka, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Cheolhong An. “OCTA-based AMD Stage Grading Enhancement via Class-Conditioned Style Transfer”, in *EMBC*, 2024. The dissertation author is the primary author of this paper. It, in part, is also a reprint of the material as it appears in Scientific Reports [6], Anna Heinke, Haochen Zhang, Krzysztof Broniarek, Katarzyna Michalska-Małecka, Wyatt Elsner, Carlo Miguel B. Galang, Daniel N. Deussen, Alexandra Warter, Fritz G. P. Kalaw, Ines D. Nagel, Akshay Agnihotri, Nehal N. Mehta, Julian Elias Klaas, Valerie Schmelter, Igor Kozak, Sally L. Baxter, Dirk-Uwe G. Bartsch, Lingyun Cheng, Cheolhong An, Truong Q. Nguyen, and William R. Freeman. “Cross-instrument optical coherence tomography-angiography (OCTA)-based prediction of age-related macular degeneration (AMD) disease activity using artificial intelligence”, *Scientific*

Reports, 2024. The dissertation author is a co-primary contributor to this work responsible for AI classifier development and experiment results analysis.

Chapter 5

Diffusion Model based Retinal Image Synthesis with Class-Condition

Building on the success of class conditioning in domain transfer tasks, we extend this approach to image synthesis. We integrate it with a state-of-the-art diffusion model to generate retinal images, thereby further expanding the training dataset. Current diffusion models operate in the VAE latent space, which supports only three-channel image encoding and decoding. Since our OCTA data consists of four projections, we first validate our method on color fundus images for Diabetic Retinopathy (DR) stage grading. We also conduct experiments on OCTA projections for AMD, as detailed in Section 5.6.

5.1 Introduction

Diabetic retinopathy (DR) is a leading cause of blindness and visual impairment, particularly among the diabetic population [80]. It is characterized by retinal lesions like microaneurysms, hemorrhages, and exudates. DR progresses through five stages: no DR, mild, moderate, severe, and proliferative DR, with increasing severity raising the risk of vision loss [81]. As prolonged

pathological states can block retinal blood vessels, early detection and grading are crucial to prevent irreversible blindness. The DR diagnostic standard has been well established on colored fundus imaging, but the shortage of ophthalmologists and the growing diabetic population strain healthcare systems. Consequently, computer-assisted tools [82, 83], including deep learning-based DR grading systems [84, 85, 86], are becoming vital to improve screening efficiency, enhance diagnostic accuracy, and ease the burden on healthcare professionals.

Recent advancements in convolutional neural networks (CNNs) have greatly improved automated DR detection. Early models like CABNet [84] used category and global attention mechanisms for better small lesion detection, while later models, such as weakly-supervised lesion-aware networks [85] and transformer-guided attention networks [86], further refined performance. However, the effectiveness of these models is heavily influenced by dataset characteristics, particularly the data distribution along different DR stages [86].

Public datasets like DDR [87] usually exhibit a severe class imbalance, with approximately half images classified as no DR, while pathological images, particularly mild, severe, and proliferative DR, are underrepresented. This imbalance leads to poor recognition performance for underrepresented categories, as observed in the confusion matrix reported in Figure 7 of [86]. Despite achieving an accuracy of 83.1%, the state-of-the-art classifier fails in real-world applications due to frequent misclassification of mild and severe DR cases. This limitation underscores the need for an alternative evaluation metric—balanced accuracy [88], which averages accuracy across all five categories. Notably, this state-of-the-art classifier achieves only 63.4% in balanced accuracy, highlighting the urgency of addressing class imbalance for reliable DR diagnosis.

Inspired by recent advancements in diffusion models [16, 89, 90] for data synthesis in computer vision tasks [91, 92, 93], we adapt these models to generate synthetic data samples to mitigate class imbalance in DR datasets. Unlike most existing image generation methods focused on visual quality, our primary aim is to improve classification accuracy. Previous research [75]

has shown a trade-off between visual and semantic quality in image restoration. Experiments in [91] also shows that it is not necessary that more realistic synthetic data is more effective for classifier training. Therefore rather than relying on visual quality oriented metrics like FID [94] or IS [95], we introduce an alternative by using the likelihood predicted by a group of pretrained classifiers as a semantic quality metric. This allows us to selectively choose synthetic samples that further benefit the classification performance. Additionally, we proposed to incorporate explicit class constraint [5] during the diffusion model’s finetuning process, guiding the model to generate samples with higher semantic relevance. Experimental results show that the semantic-oriented metric effectively filters the synthetic data and our proposed finetuning strategy achieves an improvement in balanced classification accuracy from 66.84% to 74.20%.

5.2 Related Work

5.2.1 Deep learning based DR grading

Deep learning has been extensively studied in DR grading task. Li et al. [96] applied pre-trained CNNs with transfer learning, utilizing the final fully connected layer as feature representations and employing a SVM for DR grading. Yang et al. [97] proposed a two-stage CNN that first detects lesions and then performs DR grading, achieving promising results. Several recent methods have advanced DR grading by improving class balance, lesion detection, and feature representation. CABNet [84] incorporates category and global attention blocks to address class imbalance and enhance small lesion detection. A weakly-supervised lesion-aware network [85] enhances DR grading by leveraging lesion activation maps derived from image-level supervision, improving feature discrimination without requiring pixel-level annotations. Triple-DRNet [98] enhances DR grading by distinguishing between different DR types and severity levels. CRA-Net [86] improves class-specific feature extraction and lesion relation modeling through category-relation attention mechanisms. SVPL [99] introduces a prompt-based approach that

efficiently transfers knowledge from large pre-trained models, enhancing semantic perception. But existing methods usually evaluate the performance on an imbalanced testing set with metric biased on data distribution, making the model hard to apply on real applications although achieving high overall accuracy.

5.2.2 Diffusion model and data synthesis

Diffusion models, originating from DDPM [73], have made remarkable advancements in text-to-image generation, exemplified by Stable Diffusion [89, 16], DALL-E [74], and Imagen [100]. These models leverage iterative denoising to generate high-quality visuals from textual prompts. Motivated by their generative power, researchers have explored their potential for synthetic data generation across various domains. A significant body of work focuses on utilizing synthetic data to enhance image segmentation [101, 102], while other studies investigate its impact on image classification [91, 103] and object detection [92]. The medical imaging field has also benefited from this approach. For instance, Sagers et al. [104] demonstrated that latent diffusion models can generate synthetic skin disease images to improve model performance in data-limited scenarios. Similarly, Akrouit et al. [105] employed text-to-image diffusion models to produce high-quality synthetic skin disease images, enhancing training datasets without compromising classifier performance. Furthermore, Oh et al. [106] introduced a diffusion-based data synthesis technique to address class imbalance in pathology datasets, improving both segmentation and classification outcomes. Despite these advancements, existing studies primarily address data imbalance through diffusion-based generation while overlooking the semantic quality of the synthesized samples. In contrast, our method not only generate diverse and balanced data but also ensure the semantic integrity of the synthetic samples, further enhancing their utility in downstream tasks.

5.3 Proposed Method

5.3.1 Text-to-image diffusion model finetuning on fundus images

Diffusion models are generative probabilistic frameworks that approximate a target data distribution by iteratively denoising a Gaussian-initialized sample, effectively reversing a predefined Markov chain process. In the context of text-to-image generation, a trained diffusion model f_θ generates an image $x_{\text{sync}} = f_\theta(\epsilon, c)$ from Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and a conditioning vector $c = \Phi(T)$, derived from a text prompt T via an encoder Φ . The model is trained using a mean squared error (MSE) loss to denoise a perturbed input $z_t = \alpha_t x + \sigma_t \epsilon$, where α_t, σ_t define the noise schedule over time $t \sim U([0, 1])$. These models function as a sequence of denoising networks that reconstruct clean images by leveraging reweighted variational bounds and convolution-based UNet architectures for enhanced computational efficiency. When integrated with latent spaces derived from perceptual compression techniques [16], they facilitate efficient and high-fidelity image generation by prioritizing semantically relevant information encoded in T . The text prompt T fits perfectly in our application where we can use a category-conditioned prompt: “A color fundus image in {category} stage” where $\text{category} \in \{\text{normal}, \text{mild DR}, \text{moderate DR}, \text{severe DR}, \text{proliferative DR}\}$.

Fine-tuning diffusion models requires large-scale training datasets, which are often infeasible to obtain in the medical imaging domain. To address this challenge, we adopt the diffusion model finetuning framework from DreamBooth [107] along with the LoRA finetuning strategy [108]. DreamBooth effectively trains diffusion models using only a few images (3–5) of a given subject by introducing a class-specific prior preservation loss. This loss supervises the model with additional samples belonging to the same class, ensuring that prior knowledge is retained during finetuning. This approach helps maintain diversity in the generated images while preserving class-specific characteristics, mitigating overfitting and reducing output degradation.

In our case as illustrated in top section of Figure 5.1, fundus images in the DDR

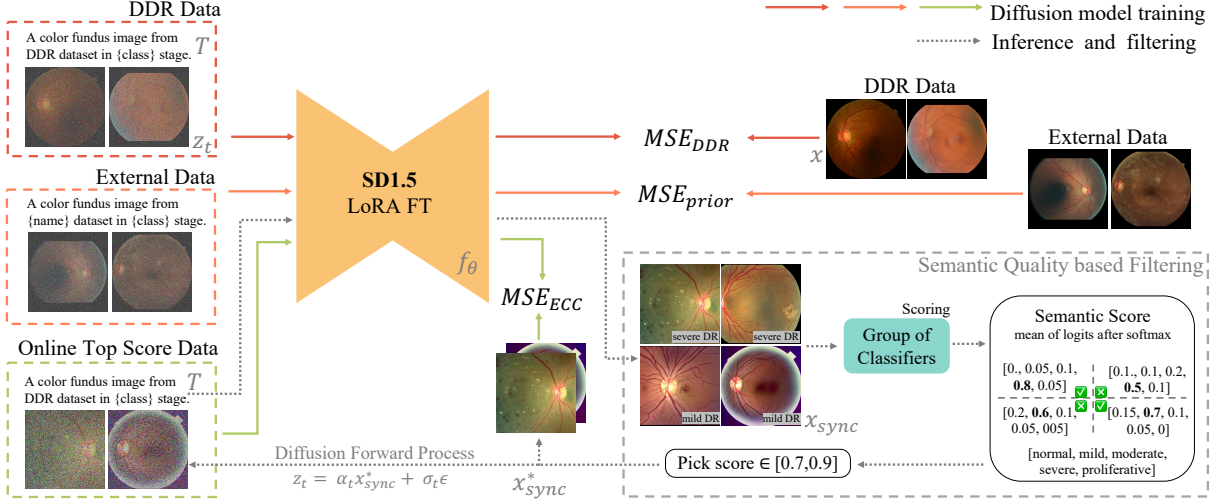


Figure 5.1: Overview of our text-to-image diffusion model finetuning framework. The top section illustrates the process of stabilizing diffusion model finetuning on fundus dataset with limited data, as described in Section 5.3.1. The bottom right depicts our proposed semantic quality evaluation and filtering pipeline, corresponding to Section 5.3.2. The entire image demonstrates how semantic quality is enhanced through explicit class conditioning during diffusion model training, as discussed in Section 5.3.3.

dataset [87] serve as subject-specific training samples MSE_{DDR} , while data for the prior-preservation term MSE_{prior} can be sourced from any retina images. Considering that fundus is the most common retina image modality, in our experiments, we incorporate fundus images from external datasets such as EyePACS¹ and APTOS² for prior-preservation loss computation. Specifically, we construct prompts in the format: “A color fundus image from {name} dataset in {category} stage”, where $\text{name} \in \{\text{DDR}, \text{EyePACS}, \text{APTOS}\}$. This integration of additional datasets ensures that the model learns particular features in DDR dataset while maintaining fundus image knowledge.

5.3.2 Semantic quality based evaluation and filtering

With the training strategy detailed in last section, we finetune a diffusion model specifically for fundus image generation within the DDR dataset domain. However as evidenced by our

¹<https://www.kaggle.com/c/diabetic-retinopathy-detection/>

²<https://www.kaggle.com/competitions/aptos2019-blindness-detection>

results in Table 5.1, it is suboptimal in enhancing classification performance by simply mixing synthetic samples with real ones to form an augmented training set. Prior research [75] has established that, in image restoration area, optimal classification performance is not achievable when images undergo restoration with distortions or perceptual losses alone. Additionally, previous experiments [91] suggest that more realistic synthetic data does not necessarily lead to better classifier training. Given these observations, we explore alternative metrics to assess the semantic quality of generated images, rather than relying solely on visual quality.

An ideal approach to measure semantic quality would involve real-time ophthalmologist feedback on each generated sample, but this is infeasible due to high costs. Inspired by prior research [109, 5] that utilized pretrained classifiers for model optimization, we propose using a group of pretrained classifiers as diagnostic agents to evaluate the usefulness of generated images for DR grading. There are no strict constraints on the architecture or training method of these classifiers, but diversity intuitively improves performance. For practical implementation, we employ the LANet [85] family, a state-of-the-art open-source DR grader for the DDR dataset, as our expert group. Since LANet has been trained on multiple CNN backbones, including VGG, DenseNet, and InceptionNet, it offers sufficient diversity to serve as a robust evaluation. We select the k models with top performance as scorers to measure the semantic quality of generated images.

With pretrained classifiers $\{g_i\}_{i=1}^k$ as scorers, we compute the semantic score for each synthetic sample x_{sync} based on its prompt-specified category j . Specifically, we extract the logits predicted by each classifier $\{\text{logit}_i\}_{i=1}^k \in \mathbb{R}^{5 \times 1}$, apply softmax normalization, compute the element-wise average and then take the j th element: $\text{score}_j = \frac{1}{k} \sum_{i=1}^k \text{softmax}(\text{logit}_i)[j]$. We then remove synthetic samples outside the range $[0.7, 0.9]$, as low-scoring samples degrade classifier performance, while perfect-scoring samples provide little new information for training.

5.3.3 Finetuning diffusion model with explicit class condition

In this section, we further explore method refines diffusion models to generate samples with high semantic quality by incorporating explicit class condition (ECC). Prior approaches [109, 5] integrate classification loss into the optimization objective. However, this approach is not suitable for our method. The reason is that we adopt the latent diffusion framework, generating images as VAE latents, for which no existing DR classifier has been trained. Thus in this section, we implement explicit class condition through an iterative self-supervised filtering and finetuning process inspired by DreamSync [110].

As illustrated as the gray dash arrows in Figure 5.1, starting from a text prompt T specifying class, the model f_θ generates a set of images $\{x_{\text{sync}}^{(k)}\}_{k=1}^N$, where $N = 256$ is empirically determined for stable finetuning. Then we utilize same evaluation strategy and filter threshold detailed in Section 5.3.2 to select samples with high semantic quality denoted as $\{x_{\text{sync}}^*\}$. High-quality image-text pairs (T, x_{sync}^*) are then iteratively used to finetune the diffusion model via a LoRA-based optimization strategy for computational efficiency. The refinement process is governed by minimizing additional reconstruction loss over the filtered top score samples (green arrow), leading to the final optimization objective:

$$L_{\text{total}} = \text{MSE}_{DDR} + \alpha \text{MSE}_{\text{prior}} + \beta \text{MSE}_{ECC}$$

where MSE_{DDR} and $\text{MSE}_{\text{prior}}$ ensure stable finetuning on limited fundus data as detailed in Section 5.3.1, while MSE_{ECC} introduces an explicit class condition to enhance semantic quality. The loss weights α and β balance the contributions of these terms for optimal performance.

5.4 Experiments for Data Imbalance Issue

5.4.1 Implement details

Datasets

In this chapter, we focus on enhancing the balanced accuracy of CNN classifiers on DDR dataset [87], while leveraging EyePACS and APTOS datasets as external sources for prior loss computation. The DDR dataset comprises of 12522 fundus images for five-class DR grading, partitioned into 6260 training images, 2503 validation images, and 3759 test images. Notably, the dataset exhibits an imbalanced class distribution. For example in the training set, the number of samples from normal to proliferative DR stages are 3133, 315, 2238, 118, and 456, respectively. We employ the proposed data synthesis to mitigate the imbalance in the training set. For the testing set, we utilize balanced accuracy to ensure that performance metrics are not biased by the class distribution.

Implementation

We implement both the diffusion model and classifiers using the PyTorch framework. The diffusion model is based on Stable Diffusion 1.5 (SD1.5) from the Hugging Face Diffusers library. Specifically, we finetune SD1.5 on 512×512 fundus images using LoRA with rank 8. To enhance data diversity, we apply center cropping and random flipping augmentations. Due to GPU memory constraints, we set the batch size to 3 but employ gradient accumulation over 32 steps to achieve a larger effective batch size. The model is trained with a constant learning rate of 1e-5 without warmup. During training, we initialize with $\beta=0$ for the first 500 steps to establish a stable starting point for the ECC term, then set $\beta=0.001$, updating the selected samples every 10 steps. For classifier training, we maintain the same preprocessing and training pipeline as described in [85]. We train Inception-v3, DenseNet-121, and VGG-16 backbones for 300, 200, and 100 epochs, respectively, using a batch size of 32. The initial learning rate is 1e-3 with

Table 5.1: Classification performance of LANet with different backbones using the proposed data synthesis on the DDR dataset. “B. Acc” is short for balanced accuracy [88]. Models marked with * are trained by the LANet authors. Basic DM and ECC DM refer to diffusion models trained as described in Sections 5.3.1 and 5.3.3, respectively. w/ filter indicates the application of the filtering strategy detailed in Section 5.3.2. We keep the total number of synthetic training samples same before and after filtering, achieved through random selection.

Backbone	VGG-16			Inception-v3			DenseNet-121		
Metric	B. Acc	Kappa	Acc	B. Acc	Kappa	Acc	B. Acc	Kappa	Acc
LANet*	65.92	86.41	83.83	67.80	85.92	83.88	64.24	84.20	82.50
LANet	66.84	85.06	83.67	65.97	84.47	81.70	62.14	84.26	83.08
+ oversampling	68.78	82.47	79.06	66.46	84.55	81.28	64.79	83.66	80.69
+ basic DM w/o filter	71.37	81.94	76.99	69.01	82.53	78.27	67.19	83.35	78.8
+ basic DM w/ filter	73.79	84.24	77.92	71.33	84.18	79.49	69.47	82.32	76.59
+ ECC DM w/o filter	74.20	81.38	75.02	71.95	83.76	76.67	70.22	83.76	79.76
+ ECC DM w/ filter	73.82	82.58	76.80	70.38	84.74	77.57	69.20	83.82	79.28

polynomial decay (power=0.9). The model with the best validation performance is selected for final evaluation.

5.4.2 Experimental results

To show the effectiveness of our proposed diffusion model training and data filtering strategy, we conduct experiment based on a state-of-the-art DR grader, LANet [85], on DDR dataset. We implement their classifiers based on author’s open-source codes and report performance of both our implementation and author’s models in Table 5.1. We also include the performance of oversampling as baseline solution for data imbalance. We focus on balanced accuracy (B. Acc) but we also provide unbalanced accuracy (Acc) and quadratic weighted kappa [85] for reference.

Main results

Table 5.1 demonstrates that oversampling enhances balanced accuracy to some extent, with a more significant improvement when using diffusion model-generated data. Among the synthetic data settings, the basic diffusion model performs the worst, underscoring the effectiveness of our proposed semantic quality-based evaluation and filtering strategy. Furthermore, finetuning

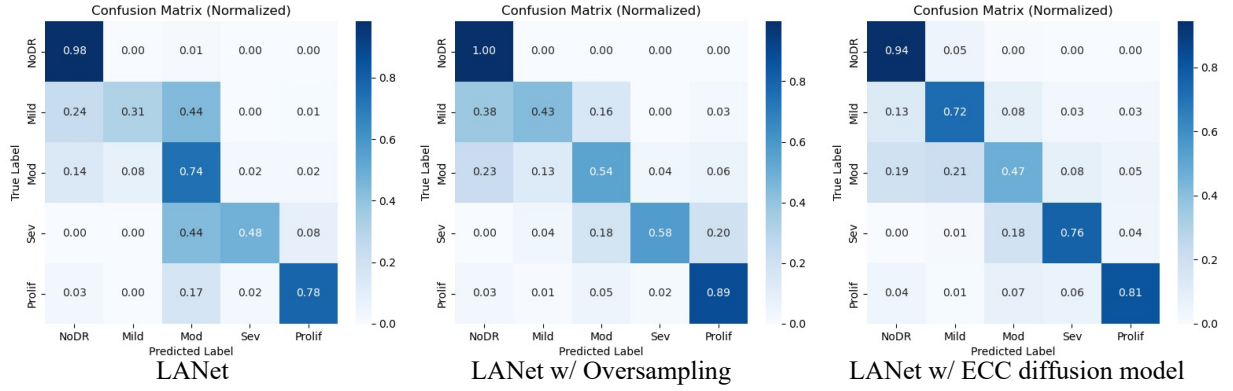


Figure 5.2: Comparison of confusion matrices without and with high-semantic-quality synthetic samples. Oversampling helps address data imbalance, while our proposed data synthesis method further enhances performance.

the diffusion model with explicit class condition enables the generation of high-semantic-quality samples, achieving similar accuracy comparable to the basic model with filtering. This validates the efficacy of our ECC finetuning strategy. An interesting observation is that with ECC, synthetic data without filtering slightly outperforms the filtered version. This suggests that a reasonable number of samples with scores outside the range $[0.7, 0.9]$ may contribute to classifier learning. Thus, an implicit semantic quality control via the diffusion model appears more beneficial than a hard threshold-based filtering approach. The confusion matrix in Figure 5.2 illustrates how synthetic data improves balanced accuracy: While this comes at the cost of reduced accuracy for moderate DR cases, the classifier exhibits improved recognition of mild and severe DR, maximizing values along the main diagonal.

Visual inspection

Figure 5.3 presents synthetic DDR samples generated by our diffusion model. In the first row, the samples on the left achieve a higher semantic score and display microaneurysms—key diagnostic features of mild DR. In contrast, the samples on the right either lack visible microaneurysms or suffer from poor image quality, making disease assessment challenging. Similarly, in the second row, the left-side samples demonstrate characteristic features of severe

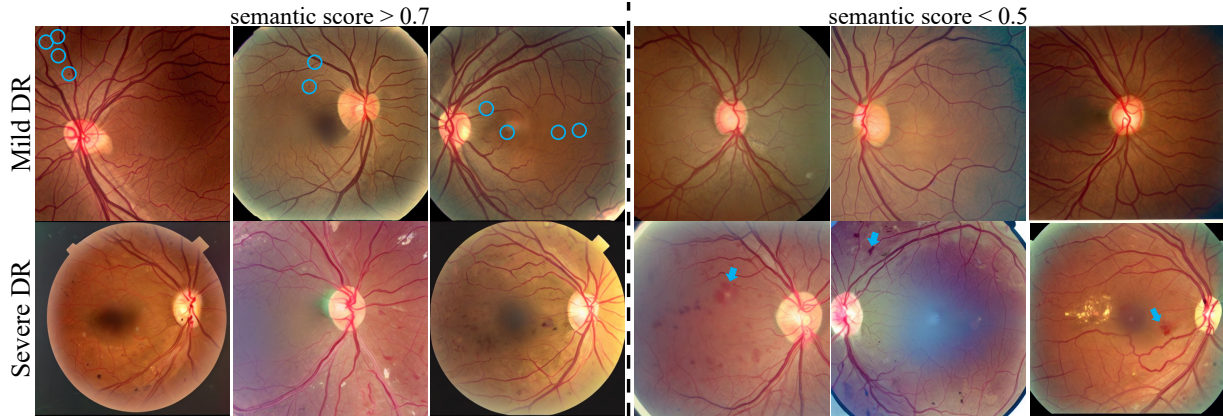


Figure 5.3: Visual inspection of synthetic samples.

non-proliferative DR, including diffuse retinal hemorrhages, microaneurysms across all four quadrants, and venous beading in at least two quadrants. In comparison, the right-side samples are more indicative of proliferative DR, as they show signs of neovascularization with intra- or subretinal hemorrhage.

5.4.3 Ablation study

Performance v.s. data amount

Figure 5.4 further examines the impact of synthetic data volume on model performance. Notably, i) while balanced accuracy improves, overall accuracy declines—a trend consistent with Table 5.1. This phenomenon aligns with expectations: the model initially learns from an imbalanced dataset, yielding high overall accuracy on a similarly skewed test set. Incorporating synthetic data for underrepresented classes mitigates this imbalance, enhancing balanced accuracy at the cost of reduced overall accuracy due to the diminishing advantage of the imbalanced class distribution. When the dataset becomes fully balanced, the two accuracy curves ultimately converge, producing an unbiased classifier. ii) The initial simultaneous increase in all three performance metrics suggests that introducing a limited amount of synthetic data into underrepresented classes does not immediately disrupt the imbalance prior of training dataset. Instead, it expands

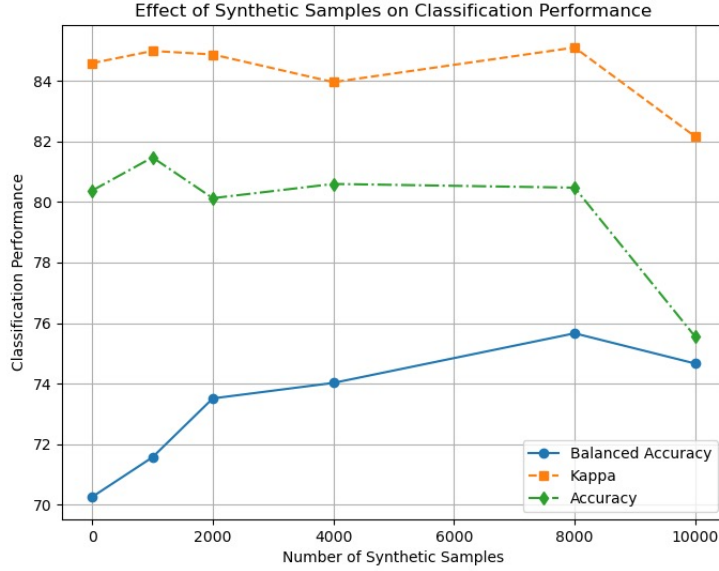


Figure 5.4: Classification performance variation w.r.t the amount of synthetic data.

the training set’s diversity while preserving key statistical properties, leading gain across all three measures. However, iii) the benefits of synthetic data diminish over time, likely due to the limited diversity of generated samples. Since real data remains fundamental to generalization, an excessive amount of synthetic data may dilute its impact, constraining further improvement.

Compared with real data

In Section 5.3.1, we propose to use external fundus dataset for prior loss calculation. One may raise questions that how about using data in those external datasets to alleviate the imbalanced issue. This section provide experiments for answers. Here we use EyePACS dataset for experiment as it has the largest amount of training samples. Specifically for underrepresented categories, mild DR and severe DR, it has 2,363 and 793 samples, respectively, after leaving 80 samples for testing. If filtered using our methods proposed in 5.3.2 with threshold $[0.65, 0.9]$, there are 753 and 112 samples left, respectively. With these data, we can form the experimnts in three setting, EyePACS all and EyePACS w/ mean using all the 3,156 unfiltered data and 865 filtered data, respectively. EyePACS w/o means we randomly select 753 from 2,363 unfiltered

Table 5.2: Performance comparison with real and synthetic data.

Metric	Balanced Acc	Kappa	Acc
LANet	66.84	85.06	83.67
+ oversampling	68.78	82.47	79.06
+ ECC DM w/o filter	74.20	81.38	75.02
+ EyePACS all	71.86	82.74	75.94
+ EyePACS w/o filter	71.38	82.32	76.68
+ EyePACS w/ filter	73.59	84.14	76.77

mild DR set and 112 from 793 unfiltered severe DR set to get a training set with same size as EyePACS w/ but without filtering.

The experimental results based on VGG backbone have been reported in Table 5.2. As observed, 1) EyePACS data helps but not optimal, especially before filtering. This might be contributed to the domain gap between DDR and EyePACS fundus datasets. This time, the samples with lower semantic score mainly represent they are out of DDR dataset distribution instead of having lower semantic quality. 2) Our diffusion model with ECC still outperforms real EyePACS data despite we use 10K synthetic samples while real data only have 3,156 samples. However, considering the cost of generating a synthetic sample and collecting a real sample, our proposed diffusion model with ECC finetuning strategy has even slightly better performance as well as much lower expense. This demonstrate the usefulness of our methods in alleviating data imbalanced issue. 3) Finally comparing the EyePACS w/o setting with EyePACS all, we can find a slight improvement by involving more EyePACS samples. However, it is not as significant as our proposed semantic based filtering strategy.

Semantic filter threshold

In this ablation study, we investigate the impact of varying the lower and upper thresholds used in the filtering strategy described in Section 5.3.2. As illustrated in Figure 5.5, the threshold pair (0.5, 0.7) yields the best performance, indicating that the model benefits from incorporating

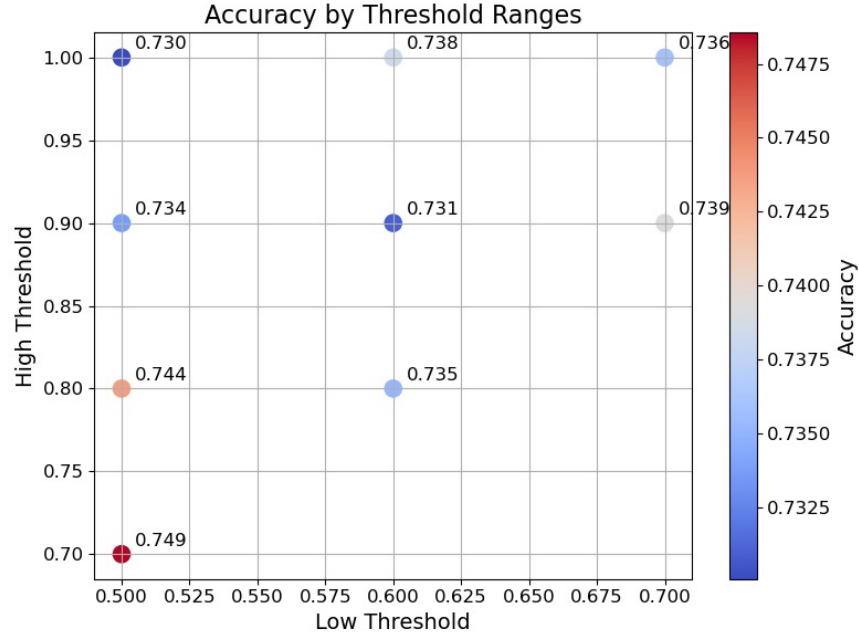


Figure 5.5: Classification performance variation w.r.t the choices of semantic score threshold.

a higher proportion of novel information for underrepresented classes rather than some typical samples. This improvement can be attributed to the fact that such samples provide more semantic diversity, enabling the classifier to learn a more discriminative decision boundary.

Oversampling weights

As shown in Figure 5.2, the classification accuracy for the moderate DR stage tends to be lower, indicating a challenge in distinguishing the three intermediate classes. To address this issue, we conduct an ablation study to examine the effect of adjusting oversampling weights during classifier training as a means to mitigate performance imbalance.

The results, presented in Figure 5.6, demonstrate that carefully tuning the oversampling weight leads to improvements. With appropriate weight adjustment, the classifier not only maintains accuracy comparable to the baseline vanilla oversampling method (without the diffusion model) but also achieves noticeable gains in correctly identifying both mild and severe DR cases. These findings highlight the classifier’s responsiveness to oversampling weight configurations

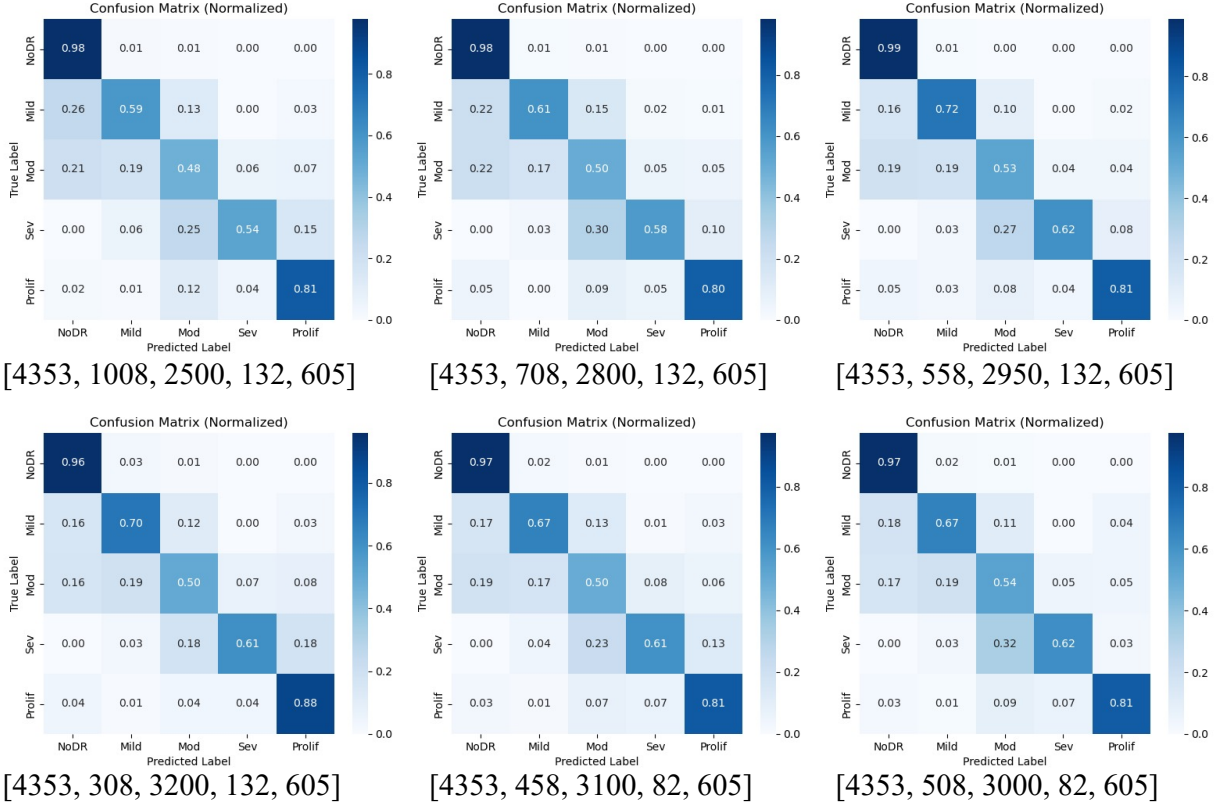


Figure 5.6: Confusion matrices corresponding to different oversampling weight configurations. We use $\{\frac{1}{n_i}\}_{i=1}^5$ as weight, so a smaller number in figure representing a larger sampling weight for this class.

and reinforce the value of incorporating flexible, class-aware sampling strategies. By doing so, the model achieves a more balanced performance, thereby improving fairness and robustness in imbalanced medical image classification tasks.

5.5 Additional Experiment for Dataset Extension

In the previous ablation study, we investigated the impact of augmenting two underrepresented categories: Mild DR and Severe DR. Since our diffusion model is capable of generating synthetic images across all five DR stages, we extend the experiment to evaluate its influence when applied to the entire dataset.

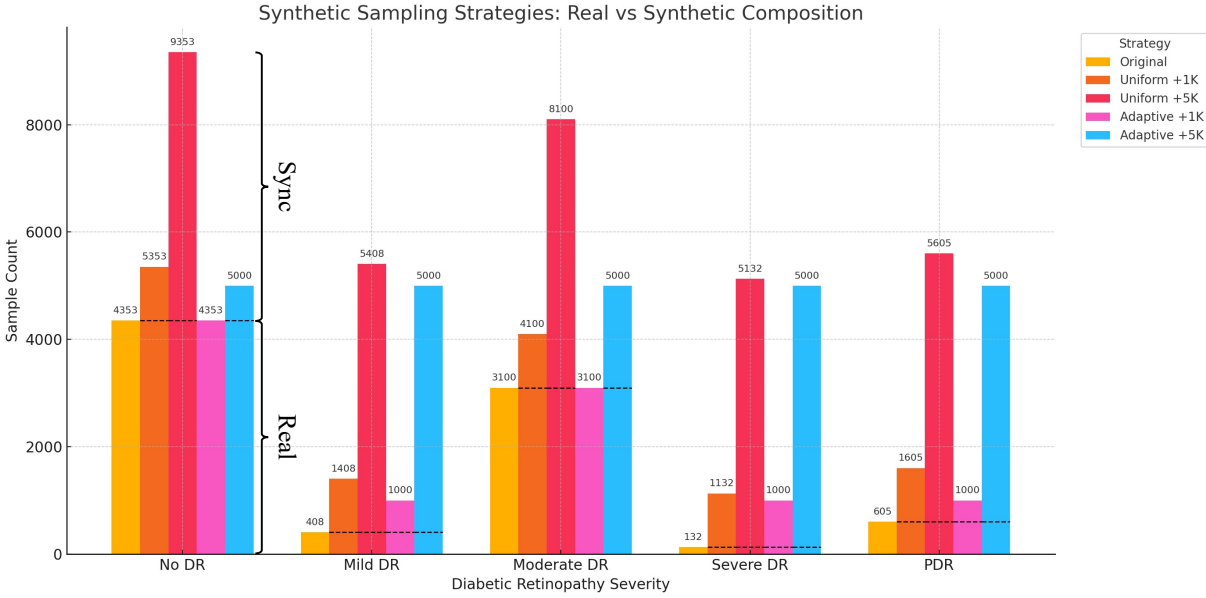


Figure 5.7: Illustration of uniform and adaptive synthetic sampling strategies. The uniform approach boosts every class equally, irrespective of class distribution. In contrast, the adaptive strategy prioritizes minority classes, such as ‘Severe DR’, by allocating a greater proportion of synthetic samples. This adaptive approach serves as a trade-off between achieving class balance (2-class setting) and extending the dataset (uniform approach).

Performance v.s. data amount

To further understand the impact of synthetic data quantity on classification performance, we conduct experiments involving all five DR classes. Given the variation in class distribution—with some classes being well-represented and others underrepresented—we compare two augmentation strategies: 1) Uniform Strategy: Synthetic samples are added equally across all classes, irrespective of their original data distribution. While simple and straightforward, this approach may overrepresent majority classes such as ‘No DR’. 2) Adaptive Strategy: This approach intends to enhance the minority classes. For each class, a minimum sample count threshold is defined. If the number of real samples for a class falls below the threshold, synthetic samples are generated to meet the minimum. Otherwise, all available real samples are used without additional augmentation. Figure 5.7 illustrates this setup. For example, to meet a 1,000-sample minimum requirement: Mild DR (408 real samples) receives 592 synthetic samples; Moderate DR (3,100

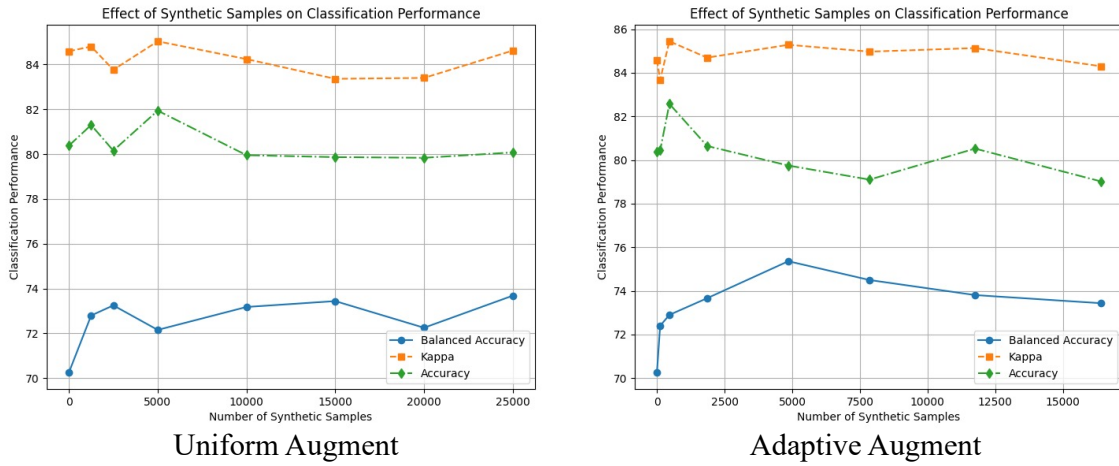


Figure 5.8: Classification performance with respect to varying amounts of synthetic data generated by uniform and adaptive sampling strategies.

real samples) requires no augmentation; Severe DR (132 real samples) is augmented with 868 synthetic samples. As a result, this classifier is trained with 1K Mild and Severe DR samples as well as 3,100 Moderate DR samples.

In our dataset, the real sample distribution across classes is as follows: 4,353 (No DR), 408 (Mild DR), 3,100 (Moderate DR), 132 (Severe DR), and 605 (Proliferative DR). Based on these counts, we determine the total number of synthetic samples needed to reach thresholds of 250, 500, 1,000, 2,000, 3,000, 4,000, and 5,000 samples per class: 118, 460, 1,855, 4,855, 7,855, 11,755, and 16,402 samples, respectively.

Figure 5.8 shows how classification performance changes with the amount of synthetic data under each sampling strategy. Compared to the previous 2-class setting (Figure 5.4), several consistent patterns emerge: i) All performance metrics improve initially with the addition of synthetic data, regardless of the sampling strategy. ii) The performance gains diminish as more synthetic data is added, highlighting diminishing returns. However in the terms of balanced accuracy, adaptive strategy has higher performance, around more than 75% with 4,855 synthetic samples, showing its ability to alleviate the data imbalanced issue. iii) Unlike the 2-class scenario, where accuracy drops with increasing numbers of synthetic samples, here the accuracy curve

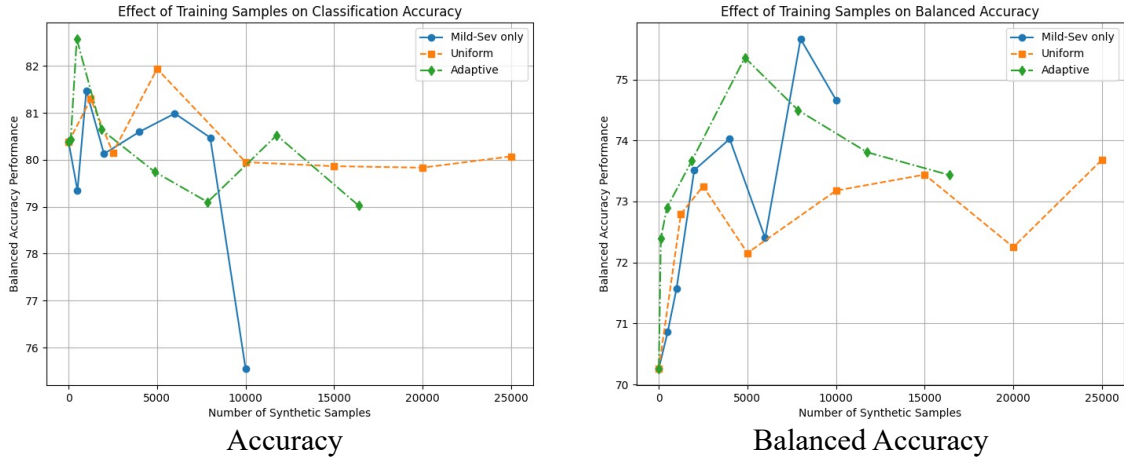


Figure 5.9: Comparison of Accuracy and Balanced Accuracy variations with increasing amounts of synthetic data under different sampling strategies.

plateaus under the uniform strategy. This is likely because uniformly augmenting all classes maintains or smoothens the original class distribution rather than drastically altering it. In contrast, the adaptive strategy causes a gradual decline in accuracy as the minimum per-class sample requirement increases. This occurs because increasing the threshold adds a disproportionate number of synthetic samples to underrepresented classes, eventually producing a fully balanced dataset (5,000 samples per class in the end). However, this accuracy drop is more moderate than in the 2-class setting, reinforcing the conception that the adaptive strategy serves as a compromise between aggressive balancing and directly extending the dataset.

To highlight this further, Figure 5.9 compares accuracy and balanced accuracy curves with respect to the total number of synthetic samples. In this view: 1) Accuracy under the uniform strategy remains stable, whereas in the 2-class setting it dropped sharply. The adaptive strategy lies between the two extremes. 2) Balanced accuracy consistently improves in all settings, with the 2-class adaptive augmentation achieving the highest peak, followed by the 5-class adaptive setting. Overall, these results confirm that synthetic samples generated by our diffusion model are effective in enhancing class balance and improving model robustness, particularly in underrepresented classes.

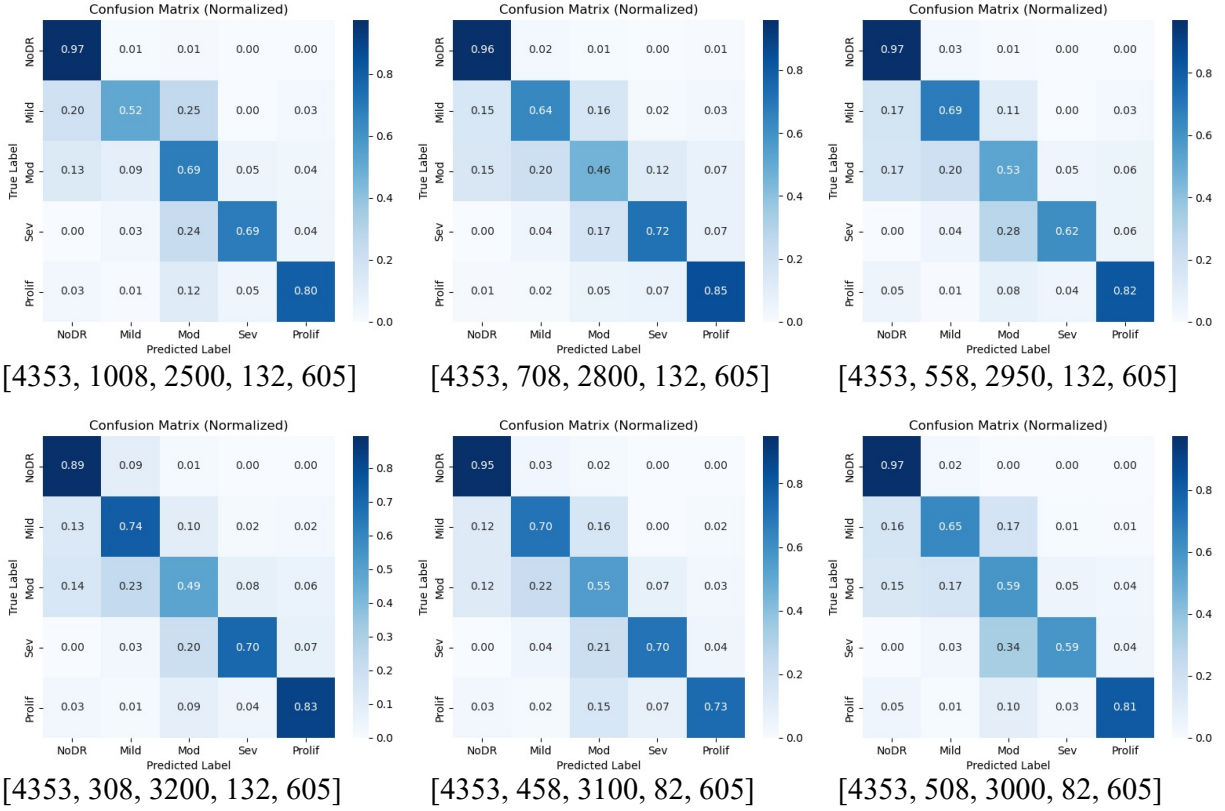


Figure 5.10: Confusion matrices corresponding to different oversampling weight configurations. We use $\{\frac{1}{n_i}\}_{i=1}^5$ as weight, so a smaller number in figure representing a larger sampling weight for this class.

Oversampling weights

We further perform an ablation study to evaluate the impact of adjusting class-wise oversampling weights during classifier training in the context of full dataset extension. The corresponding results are depicted in Figure 5.10. Consistent with the findings reported in Section 5.4.3, we observe that appropriate weight calibration allows the classifier to improve sensitivity in detecting underrepresented classes while maintaining comparable accuracy on the other classes. Moreover, the overall structure of the confusion matrices in Figure 5.10 shows a better diagonal trend and reduced misclassification rates compared to those in the 2-class setting (Figure 5.6). This improvement can be attributed to the adaptive sampling strategy, which better balances the trade-off between overall accuracy and balanced accuracy. By distributing synthetic

data more effectively across the class spectrum and pairing it with informed weight adjustments during training, the classifier is able to generalize more effectively across all DR stages, leading to more reliable and equitable performance.

5.6 Extended Results on OCTA Images for AMD

To assess the effectiveness of our proposed diffusion training methods under more challenging and realistic conditions, we applied them to the synthesis of OCTA images. OCTA is an emerging retinal imaging modality with limited publicly available datasets. Our Heidelberg OCTA dataset, used for four-stage AMD grading, is relatively small and imbalanced, comprising 471, 358, 384, and 165 samples for each stage, respectively—substantially fewer than those in the fundus DDR dataset. Owing to the scarcity of external OCTA data sources, we computed the prior loss using a diverse collection of retinal imaging modalities that capture vascular structures, including Indocyanine Green Angiography, Fluorescein Angiography, Color Fundus Photography, and Scanning Laser Ophthalmoscopy.

5.6.1 Channel problem of OCTA images

As noted in previous chapters, the 2D classifier operates on four OCTA projections, treating them as a four-channel input. However, the current latent diffusion model is limited to generating three-channel images. Ideally, this discrepancy can be addressed through one of two approaches: 1) fine-tuning the VAE components to support four-channel images, which would require a dedicated and difficult-to-acquire training dataset; or 2) interpreting the four-projection OCTA sample as a four-frame grayscale video, thereby enabling the use of a video diffusion model. While the latter approach is promising, it demands significantly greater computational resources and introduces substantial complexity into the generation pipeline. We thus leave this direction for future work.

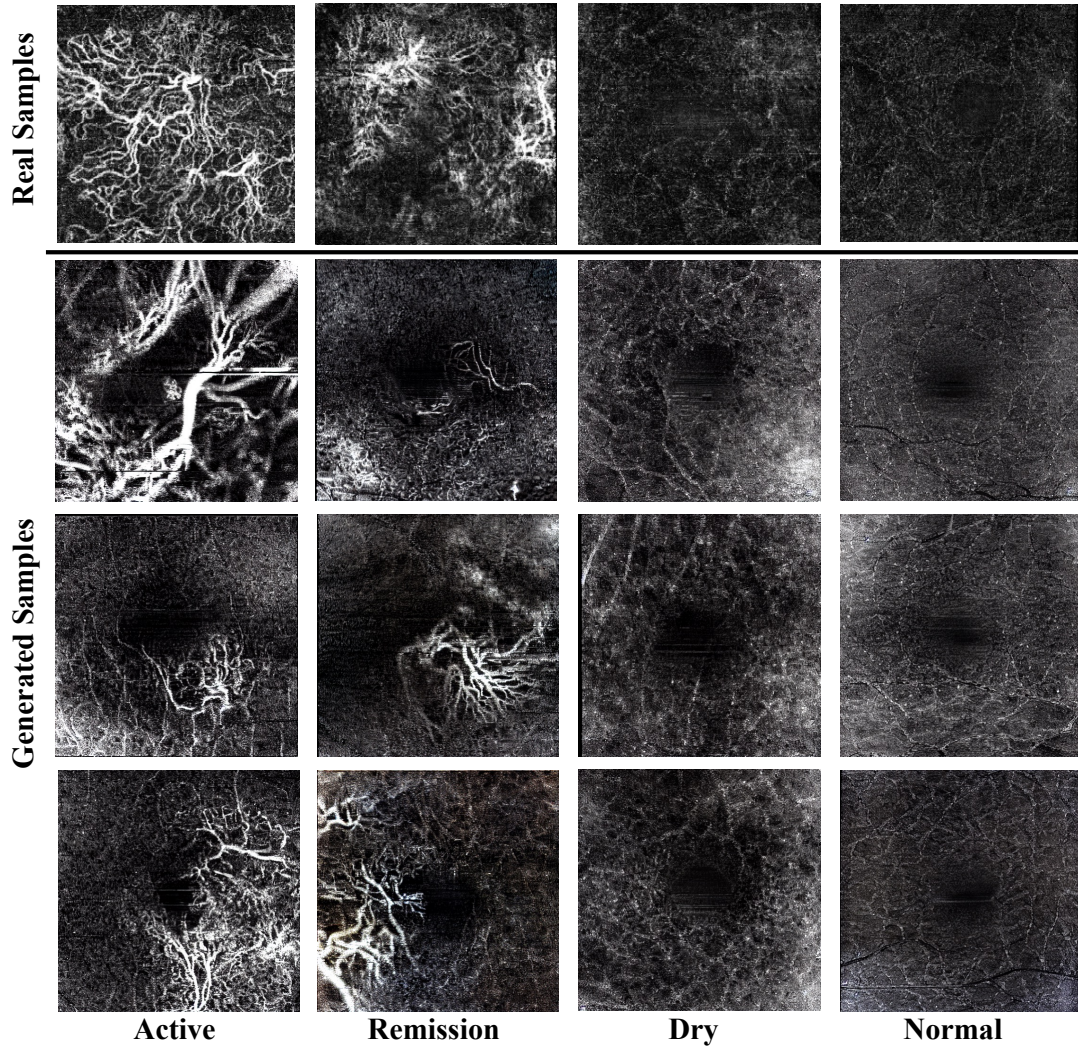


Figure 5.11: Comparison of real and generated avascular OCTA projection.

In this section, we focus on evaluating the effectiveness of our proposed diffusion model in enhancing OCTA-based AMD stage classification. Specifically, we 1) empirically demonstrate that the diffusion model is capable of synthesizing each OCTA projection individually; 2) investigate the relative importance of different OCTA projections; and 3) train classifiers and diffusion models on the most informative projection to demonstrate performance improvements.

Step 1) Figure 5.11 presents preliminary image generation results for the avascular OCTA projection across different AMD stages. These results demonstrate that the diffusion model

Table 5.3: Classification accuracy with different OCTA projections.

Projection	All four proj.	SVC only	DVC only	Ava. only	C.C. only
Accuracy	0.78	0.54	0.56	0.67	0.63

Projection	(Ava. + C.C.)/2	Three-channel	Shuffled four proj.
Accuracy	0.71	0.73	0.70

can independently synthesize each OCTA projection and capture characteristic AMD-related structural patterns to a certain extent, despite the vessel patterns appearing less realistic.

Step 2) We retrained the classifiers using different OCTA projections, following the training protocol as methods used for Table 4.3. The classification performance for each individual projection and selected combinations is summarized in Table 5.3. From the results, we observe that the avascular (Ava.) and choriocapillaris (C.C.) projections are the most informative for AMD stage grading, whereas the superficial (SVC) and deep (DVC) vascular complexes contribute less. This finding aligns with the analysis discussed in Chapter 3, Section 3.4.4.

To further probe inter-projection dependencies, we conducted a ‘shuffle’ experiment in which the avascular projection was held fixed while the remaining three projections were replaced with those from a different eye within the same stage. The resulting performance drop indicates that, although the projections vary in importance, their spatial and contextual correlation is critical for accurate AMD classification. This suggests that simply generating each projection independently and stacking them into a four-channel OCTA sample is suboptimal.

Finally, we introduced the (Ava+CC) setting, in which a grayscale image is constructed by averaging the avascular and CC projections—the two most informative channels. Additionally, we explored a three-channel configuration by averaging the SVC and DVC into one channel, and combining it with the avascular and CC projections to form a 3-channel input. Both multi-projection configurations outperformed the single-projection setting, with the 3-channel input offering the best trade-off between feasibility and classification performance. However, unlike RGB images where the channels typically exhibit similar structures, OCTA projections display

distinctly different anatomical patterns, as illustrated in Figure 5.11. This heterogeneity makes it challenging for the latent diffusion model to effectively synthesize coherent 3-channel OCTA samples. Therefore, in Step 3, we proceed with the single-projection setting and train the diffusion model to generate the averaged avascular projection.

5.6.2 Diffusion model training

Semantic score threshold

Building on the above results, we now have three key components: 1) subject-level image data from our in-house Heidelberg OCTA dataset, 2) external retinal imaging data (non-OCTA) used for prior preservation loss, and 3) a group of AMD diagnosis classifiers trained on the avascular projection. The next step involves determining the optimal threshold for selecting high semantic quality synthetic samples. To this end, we perform a series of experiments—similar in design to those presented in Figure 5.5—to identify the most suitable threshold value. However, in this case, the evaluation is conducted over a sparser range of threshold values as we already have prior expectations from Figure 5.5.

Figure 5.12 illustrates the relationship between classification accuracy and varying semantic quality thresholds. Compared to the simple oversampling baseline, the diffusion-based model demonstrates a consistent improvement in accuracy, highlighting its capacity to enhance the diversity of the training dataset. The threshold pairs of both (0.5, 0.7) and (0.7, 0.9) yield the highest performance, suggesting that the model benefits from a balanced mix of typical samples and those near the decision boundary. This behavior contrasts with trends observed in fundus image classification, potentially due to the significantly smaller size of the real OCTA training set.

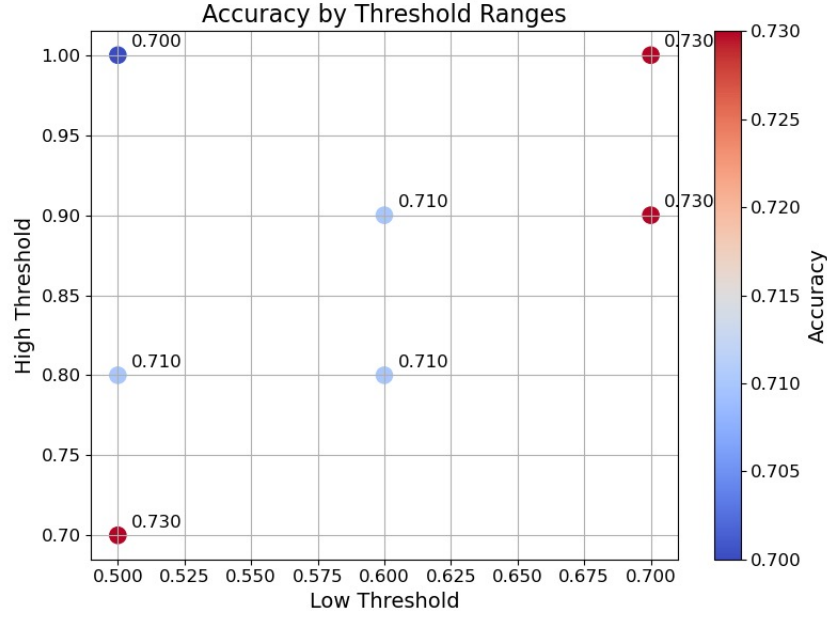


Figure 5.12: Performance w.r.t semantic quality threshold in content of Heidelberg OCTA data synthesis.

Diffusion model with class condition

In this section, we evaluate the classification performance gains enabled by incorporating synthetic OCTA data generated by our proposed diffusion models. As the Heidelberg test set is balanced with 25 samples per AMD stage, we report both accuracy and ROC-AUC in Figure 5.13. The baseline includes an oversampling strategy. Vanilla w/o F refers to a diffusion model trained without an explicit class-conditioning loss term and without filtering synthetic samples based on semantic scores. In contrast, ECC w/ F denotes a diffusion model trained with an explicit class-conditioning loss and includes semantic filtering to retain high-quality samples.

Figure 5.13 further confirms the effectiveness of diffusion-based data augmentation for AMD stage classification, as evidenced by consistent gains in both accuracy and AUC. Across all training configurations, synthetic data generated by diffusion models outperforms naive oversampling. Among the tested diffusion variants, the unfiltered vanilla model produces the lowest performance, while applying semantic filtering leads to improved results. Similarly, the

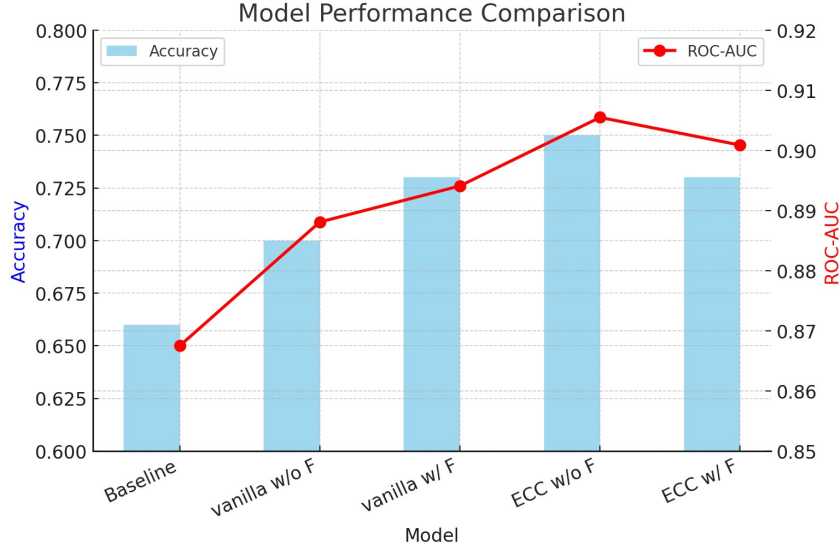


Figure 5.13: Performance w.r.t OCTA synthetic data generated by different versions of proposed diffusion model in content of Heidelberg OCTA data synthesis.

ECC-based model without filtering surpasses its filtered counterpart, indicating that soft sampling strategies may be more effective than rigid thresholding. These findings align with the trends reported in Table 5.1, underscoring the robustness and generalizability of the proposed approach in the more challenging OCTA classification setting.

5.7 Discussion

In this study, we employed a diffusion model for fundus image synthesis to address the issue of training data imbalance. Specifically, we adopted the DreamBooth framework and introduced a semantic quality metric derived from an ensemble of pretrained classifiers. By integrating both components into the training pipeline, we ensured the generation of high semantic quality fundus images within the DDR domain. Experimental results demonstrate that our method improves the balanced accuracy of state-of-the-art classifiers, thereby enhancing the practical applicability of automated DR diagnosis. Furthermore, we explored dataset expansion using synthetic images and extended our experiments to a more challenging scenario involving

Table 5.4: Per-class classification accuracy on real and synthetic validation set. The model is trained on a mixture of real and synthetic data.

	real valid set					sync valid set				
	No DR	Mild	Moderate	Severe	PDR	No DR	Mild	Moderate	Severe	PDR
2 class setting	0.98	0.71	0.63	0.55	0.76	0.0	1.0	0.05	1.0	0.21
5 class setting	0.98	0.65	0.6	0.61	0.8	1.0	1.0	0.86	1.0	1.0

OCTA image synthesis.

To identify the current bottlenecks of our diffusion model in enhancing classifier performance, we conducted an additional experiment. We trained the DR classifier, LANet, on a mixture of real and synthetic DDR samples, and evaluated validation accuracy on both a real validation set (100 samples) and a synthetic validation set (100 samples). The per-class accuracy breakdown is presented in Table 5.4.

For the 2-class setting—where synthetic samples were added only for mild and severe DR—the classifier performed well on these two classes within the synthetic validation set but showed poor performance on classes trained exclusively with real data. When synthetic samples were added for all five DR categories, the classifier achieved strong performance across all categories on the synthetic validation set. However, in both settings, the classification performance on the real validation set remained inferior to that on the synthetic one, particularly for underrepresented classes.

These results highlight a key limitation: although the diffusion model can generate an unlimited number of images, the diversity of these synthetic samples is constrained. Future work may focus on augmenting the semantic quality assessment with a diversity metric, such as clustering-based measures. Additionally, iterative co-optimization of the ECC diffusion model and the classifier group used for semantic evaluation presents a promising direction for improving both generation quality and downstream task performance.

Chapter 5, in part, has been submitted for publication of the material as it may appear in International Conference on Medical Image Computing and Computer Assisted Intervention

(MICCAI), Haochen Zhang, Anna Heinke, Ines D. Nagel, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Cheolhong An, “Class-Conditioned Image Synthesis with Diffusion for Imbalanced Diabetic Retinopathy Grading” in *MICCAI*, 2025. The dissertation author is the primary author of this submission.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Age-related Macular Degeneration (AMD) is one of the leading causes of irreversible central vision loss globally. As the disease advances from early to neovascular stages, timely detection and accurate classification become crucial for preserving vision and guiding therapeutic intervention. Optical Coherence Tomography Angiography (OCTA), a non-invasive imaging modality, provides unique capabilities in visualizing the retinal and choroidal microvasculature, particularly useful in detecting choroidal neovascularization (CNV). Despite its potential, the limited availability of OCTA data and challenges in image interpretation have hindered its widespread adoption in automated AMD diagnosis.

This dissertation investigates and advances OCTA-based AMD diagnosis using deep learning, emphasizing automated four-stage grading. It introduces the first deep learning framework for AMD stage classification based solely on OCTA data. The proposed approaches overcome key limitations by integrating 3D volumetric analysis, class-conditioned style translation across imaging devices, and synthetic data generation via diffusion models. Each component was developed to address real-world limitations in data availability, segmentation errors, and domain

shifts, culminating in a robust pipeline for accurate and generalizable AMD classification. We summarize our key contributions as follow:

- In Chapter 2, we curated a dedicated OCTA dataset tailored for AMD stage grading. This dataset comprises scans from two widely used retinal imaging instruments—Heidelberg and Optovue—establishing a solid foundation for the subsequent research presented in this thesis.
- In Chapter 3, we first identified the impact of layer segmentation errors on the performance of deep learning classifiers. To address this, we proposed two complementary technical directions: 1) direct analysis of 3D OCTA volumes to bypass the segmentation step responsible for projection errors; and 2) collection of sufficient training data to model the joint distribution of samples with and without segmentation errors. Both strategies are explored in this thesis.
- Also in Chapter 3, we developed a novel 3D classification model, enhanced by pretrained 2D backbones, and introduced a projection supervision based on the unique characteristics of OCTA data to improve shallow feature extraction. The proposed method achieved over 80% classification accuracy, significantly outperforming human experts in distinguishing subtle disease stages—particularly between “active” and “remission.” The clinical analysis was published in *RETINA* (2022) [4], and the classifier design and implementation details were presented at the International Conference on Computer Vision Workshop (2023) [3].
- In Chapter 4, we explored the second direction—dataset expansion—by integrating OCTA samples from both Heidelberg and Optovue devices to improve cross-domain classification performance and enhance model generalizability. The clinical insights from this work were published in *Scientific Reports* (2024) [6].
- Also in Chapter 4, to address domain shifts across imaging devices, we proposed a CycleGAN-based style transfer framework with explicit class-conditioning. This design

ensures that disease-specific features are preserved during style transformation. Our experiments demonstrate that incorporating class-related constraints significantly improves the quality and utility of the translated images. The methodology and implementation were presented at the Proceedings of the IEEE Engineering in Medicine and Biology Society (2024) [5].

- In Chapter 5, we extended the concept of class-conditioned modeling to image synthesis using diffusion models. We fine-tuned a latent diffusion model for OCTA image generation, incorporating semantic score-based filtering and class-conditioning to maintain the high semantic quality of generated samples. These enhancements improved classification performance in data-scarce settings. Our work on applying this approach to diabetic retinopathy (DR) classification to address data imbalance issues has been submitted to the International Conference on Medical Image Computing and Computer-Assisted Intervention (2025).

While our models demonstrate promising performance, this work also highlights several limitations. First, the generalizability of our findings may be constrained by the dataset size. Although our dataset is relatively large compared to existing OCTA resources, it may still fall short in capturing the full heterogeneity of AMD presentations across diverse populations and imaging conditions. Second, the inherent “black-box” nature of convolutional neural networks poses challenges to interpretability. In the context of medical imaging, where transparency and explainability are critical for clinical adoption, the inability to fully understand model decision-making remains a significant barrier to trust and practical deployment.

6.2 Future Work

In future work, to further advance the field of AMD diagnosis through OCTA-based approaches, the research could be further extended in the following aspects.

6.2.1 Synthetic image generation improvement

The first direction for future research involves enhancing the performance and versatility of the proposed image synthesis models. Several specific opportunities for improvement are outlined below:

- As discussed in Section 5.6, our current diffusion model struggles to generate high-quality OCTA projections across all four anatomical layers. To address this limitation, future work could explore video diffusion models, which are capable of treating the four-layer OCTA projection as a four-frame grayscale video. This approach may better preserve inter-layer coherence and improve anatomical fidelity.
- Additionally, as noted in the discussion section of Chapter 5, the diversity of the synthetic samples remains limited. Future research could incorporate diversity-aware evaluation metrics alongside semantic quality assessments to ensure the generated data are both clinically meaningful and varied, which is crucial for effective augmentation.
- Another promising avenue is the iterative co-optimization of the segmentation quality assessing classifiers and the ECC diffusion model. By jointly refining the synthesis and classification components through feedback loops, we can potentially enhance both image generation quality and the performance of downstream diagnostic tasks.
- Finally, our current model architecture is based on Stable Diffusion 1.5, which, while effective, has since been surpassed by more advanced generative frameworks. Future iterations of our pipeline could benefit from adopting newer architectures such as SDXL [89], the Diffusion Transformer [111], or recent autoregressive-based generative models [112, 113], which have shown improved performance in image quality and diversity.

6.2.2 Fine-grained classification with transformers

As shown in Table 2.2, our OCTA dataset includes nine fine-grained class labels. However, throughout this thesis, we have primarily focused on a coarser categorization into four combined classes: active, remission, dry, and normal. To enable a fine-grained AMD stage diagnosis, future work can explore several advanced methodologies.

One promising approach involves implementing a hierarchical classification framework. This strategy would first classify samples into broader categories (e.g., active, dry, and normal) and then apply specialized classifiers within each category to distinguish between finer subtypes. Such a structure could be particularly effective in leveraging the inherent semantic hierarchy of AMD progression. In this context, we can explore a range of deep learning-based fine-grained classification techniques [114], which have been successfully applied to domains where inter-class differences are subtle—such as bird species, car models, and flower varieties. Techniques such as localization and rescaling [115] and attention-based mechanisms [116] have demonstrated effectiveness in isolating and emphasizing subtle discriminative features necessary for high-resolution classification.

Moreover, Vision Transformers (ViTs) [79] have emerged as powerful alternatives to convolutional neural networks for image-based tasks. Leveraging self-attention mechanisms, ViTs excel at modeling long-range dependencies and learning robust image representations [117]. In natural image analysis, ViTs have achieved or even surpassed CNN performance across tasks such as classification, detection [118], and segmentation [119]. Their adoption in the medical imaging domain has also shown promise. For instance, Matsoukas et al. [120] experimentally verified that ViTs can match CNNs in performance with minimal modification. While CNNs may outperform ViTs when trained from scratch, ViTs pretrained on ImageNet or via self-supervised learning often deliver superior performance, especially in data-constrained scenarios. Building on this, transformer-based fine-grained classification models [121] could be employed to automatically capture intricate anatomical patterns within high-resolution OCTA projections,

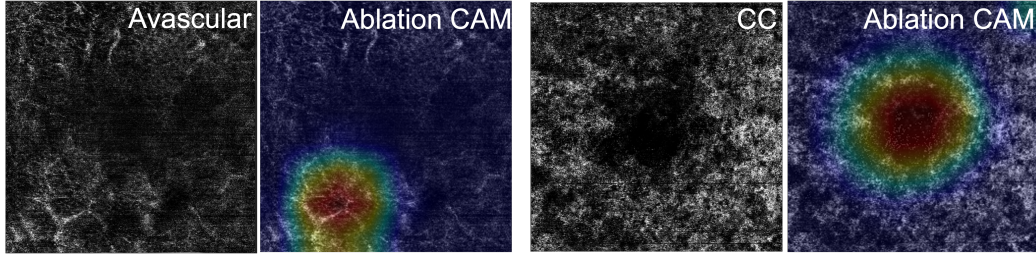


Figure 6.1: Examples of AblationCAM visualization results.

potentially improving both granularity and clinical interpretability of AMD stage classification.

6.2.3 Biomarker discovery with model visualization

Our study has demonstrated the efficacy of OCTA as a standalone modality for AMD stage grading using deep learning-based classifiers. However, a key limitation of our approach lies in the lack of interpretability—an inherent drawback of deep neural networks. The opaque, “black-box” nature of these models presents a significant challenge, particularly in clinical contexts where physicians require transparent and explainable outputs to build trust and support decision-making.

To mitigate this issue, a new line of investigation could be initiated, focusing on model interpretability through visualization techniques. These methods aim to reveal the underlying rationale behind the model’s predictions. For example, AblationCAM [122] leverages ablation analysis to determine the importance of individual feature map units with respect to a given class. It generates coarse localization maps that highlight image regions critical to the model’s decision-making.

As shown in Figure 6.1, our experiments using AblationCAM reveal insightful patterns. The generated heatmaps illustrate that the model classified the left sample as “active” due to the presence of pathological vessels in the avascular projection, while the right sample was identified as “dry” based on the dark hollows observed in the choriocapillaris layer. These explanations align with known clinical indicators, suggesting that the model may be learning biologically relevant features. However, it is important to acknowledge that the explanatory power of these

methods remains limited—particularly since even experienced ophthalmologists face difficulties achieving high diagnostic accuracy using only OCTA data.

Ultimately, we posit that meaningful progress in explainable AI for medical imaging requires a synergistic collaboration between clinicians and AI researchers. Integrating network visualization with expert clinical interpretation may not only enhance trust and adoption but also facilitate the discovery of novel, AI-driven biomarkers for disease staging.

Bibliography

- [1] S. Peralta and V. Kattouf, “Age-related macular degeneration (amd),” 2024. URL <https://www.allaboutvision.com/conditions/amd.htm> [Online; accessed May 6, 2024].
- [2] Shutterstock, “Structure of the human eye and organization of the retina. optic part of retina.” URL <https://www.shutterstock.com/image-vector/structure-human-eye-organization-retina-optic-1135203455> [Online; accessed May 6, 2024].
- [3] H. Zhang, A. Heinke, C. M. B. Galang, D. N. Deussen, B. Wen, D.-U. G. Bartsch, W. R. Freeman, T. Q. Nguyen, and C. An, “Robust amd stage grading with exclusively octa modality leveraging 3D volume,” in *ICCV Workshop*, pp. 2411–2420, 2023.
- [4] A. Heinke, H. Zhang, D. Deussen, C. M. B. Galang, A. Warter, F. G. P. Kalaw, D.-U. G. Bartsch, L. Cheng, C. An, T. Nguyen, *et al.*, “Artificial intelligence for OCTA-based disease activity prediction in age-related macular degeneration,” *RETINA*, pp. 10–1097, 2022.
- [5] H. Zhang, A. Heinke, K. B., C. M. B. Galang, D. N. Deussen, I. D. Nagel, K. M., D.-U. G. Bartsch, W. R. Freeman, T. Q. Nguyen, *et al.*, “OCTA-based amd stage grading enhancement via class-conditioned style transfer,” in *EMBC*, pp. 1–5, 2024.
- [6] A. Heinke, H. Zhang, K. Broniarek, K. Michalska-Małecka, W. Elsner, C. M. B. Galang, D. N. Deussen, A. Warter, F. Kalaw, I. Nagel, *et al.*, “Cross-instrument optical coherence tomography-angiography (OCTA)-based prediction of age-related macular degeneration (AMD) disease activity using artificial intelligence,” *Scientific Reports*, vol. 14, no. 1, p. 27085, 2024.
- [7] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, *et al.*, “Identifying medical diagnoses and treatable diseases by image-based deep learning,” *cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
- [8] R. R. Bourne, J. B. Jonas, S. R. Flaxman, J. Keeffe, J. Leasher, K. Naidoo, M. B. Parodi, K. Pesudovs, H. Price, R. A. White, *et al.*, “Prevalence and causes of vision loss in high-income countries and in eastern and central europe: 1990–2010,” *British Journal of Ophthalmology*, vol. 98, no. 5, pp. 629–638, 2014.

- [9] G. J. Coscas, M. Lupidi, F. Coscas, C. Cagini, and E. H. Souied, "Optical coherence tomography angiography versus traditional multimodal imaging in assessing the activity of exudative age-related macular degeneration: a new diagnostic challenge," *Retina*, vol. 35, no. 11, pp. 2219–2228, 2015.
- [10] L. Roisman, Q. Zhang, R. K. Wang, G. Gregori, A. Zhang, C.-L. Chen, M. K. Durbin, L. An, P. F. Stetson, G. Robbins, *et al.*, "Optical coherence tomography angiography of asymptomatic neovascularization in intermediate age-related macular degeneration," *Ophthalmology*, vol. 123, no. 6, pp. 1309–1319, 2016.
- [11] M.-L. Farecki, M. Gutfleisch, H. Faatz, K. Rothaus, B. Heimes, G. Spital, A. Lommatzsch, and D. Pauleikhoff, "Characteristics of type 1 and 2 CNV in exudative AMD in OCT-Angiography," *Graefes Archive for Clinical and Experimental Ophthalmology*, vol. 255, pp. 913–921, 2017.
- [12] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, pp. 1026–1034, 2015.
- [13] Y. Ma, H. Hao, J. Xie, H. Fu, J. Zhang, J. Yang, Z. Wang, J. Liu, Y. Zheng, and Y. Zhao, "ROSE: A retinal oct-angiography vessel segmentation dataset and new model," *IEEE transactions on medical imaging*, vol. 40, no. 3, pp. 928–939, 2020.
- [14] M. Li, Y. Zhang, Z. Ji, K. Xie, S. Yuan, Q. Liu, and Q. Chen, "IPN-v2 and OCTA-500: Methodology and dataset for retinal image segmentation," *arXiv preprint arXiv:2012.07261*, 2020.
- [15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, pp. 2223–2232, 2017.
- [16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, pp. 10684–10695, 2022.
- [17] T. D. Keenan, C. A. Cukras, and E. Y. Chew, "Age-related macular degeneration: epidemiology and clinical aspects," *Age-related Macular Degeneration: From Clinic to Genes and Back to Patient Management*, pp. 1–31, 2021.
- [18] P. Mitchell, G. Liew, B. Gopinath, and T. Y. Wong, "Age-related macular degeneration," *The Lancet*, vol. 392, no. 10153, pp. 1147–1159, 2018.
- [19] M. J. Heiferman and A. A. Fawzi, "Progression of subclinical choroidal neovascularization in age-related macular degeneration," *PLoS One*, vol. 14, no. 6, p. e0217805, 2019.
- [20] A. Heinke, W. R. Freeman, D.-U. G. Bartsch, L. Cheng, C. M. B. Galang, A. Warter, F. Kalaw, H. Zhang, T. Q. Nguyen, and C. An, "Quantitative evaluation of morphological changes in anti-vegf treated choroidal neovascularization due to age related macular degeneration using optical coherence tomography angiography," *Investigative Ophthalmology & Visual Science*, vol. 63, no. 7, pp. 1348–F0182, 2022.

- [21] F. Hajizadeh, R. Kafieh, and M. Tajmirriahi, "Introduction to optical coherence tomography," in *Atlas of ocular optical coherence tomography*, pp. 1–34, 2023.
- [22] C. V. Regatieri, L. Branchini, and J. S. Duker, "The role of spectral-domain oct in the diagnosis and management of neovascular age-related macular degeneration," *Ophthalmic Surgery, Lasers and Imaging Retina*, vol. 42, no. 4, pp. S56–S66, 2011.
- [23] V. Manjunath, J. Goren, J. G. Fujimoto, and J. S. Duker, "Analysis of choroidal thickness in age-related macular degeneration using spectral-domain optical coherence tomography," *American journal of ophthalmology*, vol. 152, no. 4, pp. 663–668, 2011.
- [24] G. Querques, L. Querques, R. Forte, N. Massamba, F. Coscas, and E. H. Souied, "Choroidal changes associated with reticular pseudodrusen," *Investigative ophthalmology & visual science*, vol. 53, no. 3, pp. 1258–1263, 2012.
- [25] R. F. Spaide, J. G. Fujimoto, N. K. Waheed, S. R. Sadda, and G. Staurengi, "Optical coherence tomography angiography," *Progress in retinal and eye research*, vol. 64, pp. 1–55, 2018.
- [26] E. W. Schneider and S. C. Fowler, "Optical coherence tomography angiography in the management of age-related macular degeneration," *Current opinion in ophthalmology*, vol. 29, no. 3, pp. 217–225, 2018.
- [27] J. Ma, R. Desai, P. Nesper, M. Gill, A. Fawzi, and D. Skondra, "Optical coherence tomographic angiography imaging in age-related macular degeneration," *Ophthalmology and eye diseases*, vol. 9, p. 1179172116686075, 2017.
- [28] J. Wu, H. Fang, F. Li, H. Fu, F. Lin, J. Li, L. Huang, Q. Yu, S. Song, X. Xu, *et al.*, "Gamma challenge: Glaucoma grading from multi-modality images," *arXiv preprint arXiv:2202.06511*, 2022.
- [29] H. Fang, F. Li, H. Fu, X. Sun, X. Cao, F. Lin, J. Son, S. Kim, G. Quellec, S. Matta, *et al.*, "ADAM challenge: Detecting age-related macular degeneration from fundus images," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2828–2847, 2022.
- [30] M. Díaz, J. Novo, M. G. Penedo, and M. Ortega, "Automatic extraction of vascularity measurements using OCT-A images," *Procedia Computer Science*, vol. 126, pp. 273–281, 2018.
- [31] M. Alam, D. Le, T. Son, J. I. Lim, and X. Yao, "AV-Net: deep learning for fully automated artery-vein classification in optical coherence tomography angiography," *Biomedical optics express*, vol. 11, no. 9, pp. 5249–5257, 2020.
- [32] Y. Guo, A. Camino, J. Wang, D. Huang, T. S. Hwang, and Y. Jia, "MEDnet, a neural network for automated detection of avascular area in OCT angiography," *Biomedical optics express*, vol. 9, no. 11, pp. 5147–5158, 2018.

- [33] M. Li, W. Zhang, and Q. Chen, "Image magnification network for vessel segmentation in OCTA images," in *Chinese Conference on Pattern Recognition and Computer Vision*, pp. 426–435, 2022.
- [34] M. Li, Y. Chen, Z. Ji, K. Xie, S. Yuan, Q. Chen, and S. Li, "Image projection network: 3D to 2D image segmentation in OCTA images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3343–3354, 2020.
- [35] D. Le, M. Alam, C. K. Yao, J. I. Lim, Y.-T. Hsieh, R. V. Chan, D. Toslak, and X. Yao, "Transfer learning for automated OCTA detection of diabetic retinopathy," *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 35–35, 2020.
- [36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, pp. 248–255, 2009.
- [37] L. Lin, Z. Wang, J. Wu, Y. Huang, J. Lyu, P. Cheng, J. Wu, and X. Tang, "BSDA-net: A boundary shape and distance aware joint learning framework for segmenting and classifying OCTA images," in *MICCAI*, pp. 65–75, 2021.
- [38] S. Yu, Y. Liu, J. Zhang, J. Xie, Y. Zheng, J. Liu, and Y. Zhao, "Cross-domain depth estimation network for 3D vessel reconstruction in OCT angiography," in *MICCAI*, pp. 13–23, 2021.
- [39] Y. Wang, Y. Shen, M. Yuan, J. Xu, B. Yang, C. Liu, W. Cai, W. Cheng, and W. Wang, "A deep learning-based quality assessment and segmentation system with a large-scale benchmark dataset for optical coherence tomographic angiography image," *arXiv preprint arXiv:2107.10476*, 2021.
- [40] Y. Zhang, C. Huang, M. Li, S. Xie, K. Xie, Z. Ji, S. Yuan, and Q. Chen, "Robust layer segmentation against complex retinal abnormalities for en face OCTA generation," in *MICCAI*, pp. 647–655, 2020.
- [41] A. M. Alqudah, "AOCT-NET: A convolutional network automated classification of multi-class retinal diseases using spectral-domain optical coherence tomography images," *Medical & biological engineering & computing*, vol. 58, no. 1, pp. 41–53, 2020.
- [42] N. Motozawa, G. An, S. Takagi, S. Kitahata, M. Mandai, Y. Hirami, H. Yokota, M. Akiba, A. Tsujikawa, M. Takahashi, *et al.*, "Optical coherence tomography-based deep-learning models for classifying normal and age-related macular degeneration and exudative and non-exudative age-related macular degeneration changes," *Ophthalmology and therapy*, vol. 8, pp. 527–539, 2019.
- [43] V. Das, S. Dandapat, and P. K. Bora, "Multi-scale deep feature fusion for automated classification of macular pathologies from OCT images," *Biomedical signal processing and Control*, vol. 54, p. 101605, 2019.

- [44] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, "Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification," *TMI*, vol. 38, no. 8, pp. 1959–1970, 2019.
- [45] X. He, L. Fang, H. Rabbani, X. Chen, and Z. Liu, "Retinal optical coherence tomography image classification with label smoothing generative adversarial network," *Neurocomputing*, vol. 405, pp. 37–47, 2020.
- [46] K. Alsaih, M. Z. Yusoff, T. B. Tang, I. Faye, and F. Mériaudeau, "Deep learning architectures analysis for age-related macular degeneration segmentation on optical coherence tomography scans," *Computer methods and programs in biomedicine*, vol. 195, p. 105566, 2020.
- [47] N. Rakocz, J. N. Chiang, M. G. Nittala, G. Corradetti, L. Tiosano, S. Velaga, M. Thompson, B. L. Hill, S. Sankararaman, J. L. Haines, *et al.*, "Automated identification of clinical features from sparsely annotated 3-dimensional medical imaging," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–13, 2021.
- [48] D. B. Russakoff, A. Lamin, J. D. Oakley, A. M. Dubis, and S. Sivaprasad, "Deep learning for prediction of AMD progression: A pilot study," *Investigative ophthalmology & visual science*, vol. 60, no. 2, pp. 712–722, 2019.
- [49] E. Vaghefi, S. Hill, H. M. Kersten, and D. Squirrell, "Multimodal retinal image analysis via deep learning for the diagnosis of intermediate dry age-related macular degeneration: A feasibility study," *Journal of Ophthalmology*, vol. 2020, 2020.
- [50] K. Jin, Y. Yan, M. Chen, J. Wang, X. Pan, X. Liu, M. Liu, L. Lou, Y. Wang, and J. Ye, "Multimodal deep learning with feature level fusion for identification of choroidal neovascularization activity in age-related macular degeneration," *Acta Ophthalmologica*, vol. 100, no. 2, pp. e512–e520, 2022.
- [51] K. A. Thakoor, J. Yao, D. Bordbar, O. Moussa, W. Lin, P. Sajda, and R. W. Chen, "A multimodal deep learning system to distinguish late stages of AMD and to compare expert vs. AI ocular biomarkers," *Scientific reports*, vol. 12, no. 1, pp. 1–11, 2022.
- [52] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, pp. 6105–6114, 2019.
- [53] R. Rocholz, M. M. Teussink, R. Dolz-Marco, C. Holzhey, J. F. Dechent, A. Tafreshi, and S. Schulz, "SPECTRALIS Optical Coherence Tomography Angiography (OCTA): Principles and Clinical Applications." from https://www.heidelbergengineering.com/media/e-learning/Totara/Dateien/pdf-tutorials/210111-001_SPECTRALIS%20OCTA%20-%20Principles%20and%20Clinical%20Applications_EN.pdf.
- [54] J. Schottenhamml, E. M. Moulton, S. B. Ploner, S. Chen, E. Novais, L. Husvogt, J. S. Duker, N. K. Waheed, J. G. Fujimoto, and A. K. Maier, "OCT-OCTA segmentation: Combining

- structural and blood flow information to segment Bruch's membrane," *Biomedical Optics Express*, vol. 12, no. 1, pp. 84–99, 2021.
- [55] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, pp. 234–241, 2015.
 - [56] B. Barz and J. Denzler, "Deep learning on small datasets without pre-training using cosine loss," in *WACV*, pp. 1371–1380, 2020.
 - [57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.
 - [58] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer learning for 3D medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.
 - [59] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*, pp. 20–36, 2016.
 - [60] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *ICCV*, pp. 5533–5541, 2017.
 - [61] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *CVPR*, pp. 6299–6308, 2017.
 - [62] M. Arya, A. S. Sabrosa, J. S. Duker, and N. K. Waheed, "Choriocapillaris changes in dry age-related macular degeneration and geographic atrophy: A review," *Eye and Vision*, vol. 5, no. 1, pp. 1–7, 2018.
 - [63] M. Cavichini, K. C. Dans, M. Jhingan, M. J. Amador-Patarroyo, S. Borooah, D.-U. Bartsch, E. Nudleman, and W. R. Freeman, "Evaluation of the clinical utility of optical coherence tomography angiography in age-related macular degeneration," *British Journal of Ophthalmology*, vol. 105, no. 7, pp. 983–988, 2021.
 - [64] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *transactions on visualization and computer graphics*, vol. 26, no. 11, pp. 3365–3385, 2019.
 - [65] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *CVPR*, pp. 2414–2423, 2016.
 - [66] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, "Controlling perceptual factors in neural style transfer," in *CVPR*, pp. 3985–3993, 2017.
 - [67] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," in *IJCAI*, 2017.
 - [68] S. Li, X. Xu, L. Nie, and T.-S. Chua, "Laplacian-steered neural style transfer," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1716–1724, 2017.

- [69] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, “Texture networks: Feed-forward synthesis of textures and stylized images,” in *ICML*, pp. 2027–2041, 2016.
- [70] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *ECCV*, pp. 694–711, 2016.
- [71] C. Li and M. Wand, “Precomputed real-time texture synthesis with markovian generative adversarial networks,” in *ECCV*, pp. 702–716, 2016.
- [72] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis,” in *CVPR*, pp. 6924–6932, 2017.
- [73] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *NeurIPS*, vol. 33, pp. 6840–6851, 2020.
- [74] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.
- [75] D. Liu, H. Zhang, and Z. Xiong, “On the classification-distortion-perception tradeoff,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 1–10, 2019.
- [76] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [77] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in *ICCV*, pp. 2794–2802, 2017.
- [78] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” in *CVPR*, pp. 11976–11986, 2022.
- [79] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [80] N. H. Cho, J. E. Shaw, S. K., Y. Huang, J. D. da R. F., A. Ohlrogge, and B. Malanda, “IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,” *Diabetes research and clinical practice*, vol. 138, pp. 271–281, 2018.
- [81] C. P. Wilkinson, F. L. Ferris III, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, J. T. Verdaguer, *et al.*, “Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales,” *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, 2003.

- [82] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *jama*, vol. 316, no. 22, pp. 2402–2410, 2016.
- [83] E. Pekel Özmen and T. Özcan, “Diagnosis of diabetes mellitus using artificial neural network and classification and regression tree optimized with genetic algorithm,” *Journal of Forecasting*, vol. 39, no. 4, pp. 661–670, 2020.
- [84] A. He, T. Li, N. Li, K. Wang, and H. Fu, “Cabnet: Category attention block for imbalanced diabetic retinopathy grading,” *TMI*, vol. 40, no. 1, pp. 143–153, 2020.
- [85] J. Hou, F. X., J. Xu, R. Feng, Y. Z., H. Zou, L. Lu, and W. Xue, “Diabetic retinopathy grading with weakly-supervised lesion priors,” in *ICASSP*, pp. 1–5, 2023.
- [86] F. Zang and H. Ma, “CRA-Net: Transformer guided category-relation attention network for DR grading,” *Computers in Biology and Medicine*, vol. 170, p. 107993, 2024.
- [87] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. K., “Diagnostic assessment of deep learning algorithms for DR screening,” *Information Sciences*, vol. 501, pp. 511 – 522, 2019.
- [88] M. Grandini, E. Bagli, and G. Visani, “Metrics for multi-class classification: an overview,” *arXiv preprint arXiv:2008.05756*, 2020.
- [89] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, “SDXL: Improving latent diffusion models for high-resolution image synthesis,” *arXiv preprint arXiv:2307.01952*, 2023.
- [90] S. Xiao, Y. Wang, J. Zhou, H. Yuan, X. Xing, R. Yan, S. Wang, T. Huang, and Z. Liu, “OmniGen: Unified image generation,” *arXiv preprint arXiv:2409.11340*, 2024.
- [91] M. Ye-Bin, N. Hyeon-Woo, W. Choi, N. Kim, S. Kwak, and T.-H. Oh, “Exploiting synthetic data for data imbalance problems: Baselines from a data perspective,” *arXiv preprint arXiv:2308.00994*, 2023.
- [92] R. Voetman, M. Aghaei, and K. Dijkstra, “The big data myth: Using diffusion models for dataset generation to train deep detection models,” *arXiv preprint arXiv:2306.09762*, 2023.
- [93] X. Yu, G. Li, W. Lou, S. Liu, X. Wan, Y. Chen, and H. Li, “Diffusion-based data augmentation for nuclei image segmentation,” in *MICCAI*, pp. 592–602, 2023.
- [94] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” *NeurIPS*, vol. 30, 2017.
- [95] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” *NeurIPS*, vol. 29, 2016.

- [96] X. Li, T. Pang, B. Xiong, W. Liu, P. Liang, and T. Wang, “Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification,” in *CISP-BMEI*, pp. 1–11, 2017.
- [97] Y. Yang, T. Li, W. Li, H. Wu, W. Fan, and W. Zhang, “Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks,” in *MICCAI*, pp. 533–540, 2017.
- [98] M. Jian, H. Chen, C. Tao, X. Li, and G. Wang, “Triple-DRNet: A triple-cascade convolution neural network for diabetic retinopathy grading using fundus images,” *Computers in Biology and Medicine*, vol. 155, p. 106631, 2023.
- [99] Y. Zhang, X. Ma, K. Huang, M. Li, and P.-A. Heng, “Semantic-oriented visual prompt learning for diabetic retinopathy grading on fundus images,” *TMI*, 2024.
- [100] J. Baldridge, J. Bauer, M. Bhutani, N. Brichtova, A. Bunner, K. Chan, Y. Chen, S. Dieleman, Y. Du, Z. Eaton-Rosen, *et al.*, “Imagen 3,” *arXiv preprint arXiv:2408.07009*, 2024.
- [101] Q. Nguyen, T. Vu, A. Tran, and K. Nguyen, “Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation,” *NeurIPS*, vol. 36, pp. 76872–76892, 2023.
- [102] J. Xie, W. Li, X. Li, Z. Liu, and C. C. Loy, “Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation,” *IJCV*, pp. 1–20, 2024.
- [103] S. Azizi, S. Kornblith, C. Saharia, and D. J. Fleet, “Synthetic data from diffusion models improves imagenet classification,” *arXiv preprint arXiv:2304.08466*, 2023.
- [104] L. W. Sagers, J. A. Diao, M. Luke, M. Groh, P. Rajpurkar, A. S. Adamson, V. R., R. Daneshjou, and A. K. Manrai, “Augmenting medical image classifiers with synthetic data from latent diffusion models,” *arXiv preprint arXiv:2308.12453*, 2023.
- [105] M. Akrou, B. Gyepesi, P. Holló, A. Poór, B. Kincső, S. Solis, K. Cirone, J. Kawahara, D. Slade, L. Abid, *et al.*, “Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images,” in *MICCAI*, pp. 99–109, 2023.
- [106] H.-J. Oh and W.-K. Jeong, “Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets,” in *MICCAI*, pp. 337–345, 2023.
- [107] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation,” in *CVPR*, pp. 22500–22510, 2023.
- [108] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, “LoRA: Low-rank adaptation of large language models,” in *ICLR*, 2021.

- [109] Z. Liu, H. Wang, T. Zhou, Z. Shen, B. Kang, E. Shelhamer, and T. Darrell, “Exploring simple and transferable recognition-aware image processing,” *TPAMI*, vol. 45, no. 3, pp. 3032–3046, 2022.
- [110] J. Sun, D. Fu, Y. Hu, S. Wang, R. Rassin, D.-C. Juan, D. Alon, C. Herrmann, S. van Steenkiste, R. Krishna, *et al.*, “Dreamsync: Aligning text-to-image generation with image understanding feedback,” in *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2023.
- [111] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *ICCV*, pp. 4195–4205, 2023.
- [112] P. Sun, Y. Jiang, S. Chen, S. Zhang, B. Peng, P. Luo, and Z. Yuan, “Autoregressive model beats diffusion: Llama for scalable image generation,” *arXiv preprint arXiv:2406.06525*, 2024.
- [113] D. Liu, S. Zhao, L. Zhuo, W. Lin, Y. Qiao, H. Li, and P. Gao, “Lumina-mGPT: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining,” *arXiv preprint arXiv:2408.02657*, 2024.
- [114] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, “Fine-grained image analysis with deep learning: A survey,” *TPAMI*, vol. 44, no. 12, pp. 8927–8948, 2021.
- [115] S. Cai, W. Zuo, and L. Zhang, “Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization,” in *ICCV*, pp. 511–520, 2017.
- [116] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, “Dual cross-attention learning for fine-grained visual categorization and object re-identification,” in *CVPR*, pp. 4692–4702, 2022.
- [117] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, “Transformers in medical imaging: A survey,” *Medical Image Analysis*, vol. 88, p. 102802, 2023.
- [118] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *ECCV*, pp. 213–229, 2020.
- [119] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding transfer learning for medical imaging,” *NeurIPS*, vol. 32, 2019.
- [120] C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, “Is it time to replace cnns with transformers for medical images?,” *arXiv preprint arXiv:2108.09038*, 2021.
- [121] Q. Xu, J. Wang, B. Jiang, and B. Luo, “Fine-grained visual classification via internal ensemble learning transformer,” *TMM*, vol. 25, pp. 9015–9028, 2023.

- [122] H. G. Ramaswamy *et al.*, “Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization,” in *WACV*, pp. 983–991, 2020.