ADVANCING RETINAL VESSEL RECOGNITION VIA DEEP LEARNING WITH LIMITED ANNOTATION

By

Dewei Hu

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Electrical and Computer Engineering

August 9, 2024

Nashville, Tennessee

Approved:

Ipek Oguz, Ph.D.

Yuankai Tao, Ph.D.

Benoit Dawant, Ph.D.

Jack Noble, Ph.D.

Yuankai Huo, Ph.D.

ACKNOWLEDGMENTS

My work is supported by the NIH grant R01EY033969, R01EY030490, R01EY031769 and the Vanderbilt University Discovery Grant Program. Other than the funding, I have obtained substantial support from a lot of people along the way to completing my doctoral degree. Foremost, I owe a debt of gratitude to my advisor, Dr. Ipek Oguz, from whom I learnt the essential methodology of approaching a research problem. She is always open to all the possibilities in my work and provide insightful suggestions. Her encouragement to pursue innovative approaches have truly shaped all my works. I will always remember that she stay up until midnight to help me generate the tractography on my vessel maps.

Next, I want to express my sincere thanks to Dr.Yuankai Tao for providing the crucial data that formed the foundation of my research. More importantly, I really appreciate his critical perspective from a biomedical engineer. This is particularly valuable since the AI solutions need to be very solid and persuasive before people from different backgrounds are willing to use.

I would also like to extend my heartfelt thanks to Dr.Benoit Dawant, Dr.Jack Noble and Dr.Yuankai Huo. The classes they taught during my first two years at the Vanderbilt University played a significant role in shaping my basic understanding of the medical image analysis area and greatly influenced my academic journey.

TABLE OF CONTENTS

Page

LIST OF TABLES . . . . **vi**

LIST OF FIGURES . . . . **vii**

1 Introduction and Specific Aims **1**

1.1 Background 1
1.1.1 Significance of analysing retinal vessels 1
1.1.2 Deep learning approach for medical image analysis 2
1.1.3 Major ophthalmic imaging modalities 3
1.1.3.1 Fundus photography 3
1.1.3.2 Fluorescein angiography 5
1.1.3.3 Optical coherence tomography 5
1.1.4 Challenges 7
1.1.4.1 Low image quality 7
1.1.4.2 Insufficient manual annotations 8
1.1.4.3 Domain shift 8
1.2 Specific Aims 9
1.2.1 OCT image denoising 9
1.2.2 Retinal vessel enhancement 9
1.2.3 Domain adaptation and generalization 9
1.3 Overview of the thesis 10

2 OCT Denoising **11**

2.1 Introduction 11
2.2 Related Work 11
2.2.1 Supervised methods 12
2.2.1.1 Direct prediction 12
2.2.1.2 Residual prediction 12
2.2.2 Unsupervised methods 12
2.2.2.1 Unpaired data 12
2.2.2.2 Noise2Noise 13
2.3 Datasets 13
2.4 Evaluation metrics 14
2.5 Preliminary: Self-fusion 15
2.6 Supervised-learning Method: PMFN 18
2.6.1 Method 18
2.6.1.1 Network I: Pseudo-modality creation 18
2.6.1.2 Network II: Pseudo-multimodal fusion 19
2.6.2 Experiments 19
2.6.3 Results 20
2.6.3.1 Qualitative results 20
2.6.3.2 Quantitative results 23
2.7 Unsupervised-learning Method: DDPM 25
2.7.1 Method 25
2.7.2 Experiments 27
2.7.3 Results 28

2.8 Chapter summary 31

3 Enhanced Angiogram Synthesis 32

3.1 Introduction 32
3.2 Datasets 33
3.3 Preliminary: Vessel Enhancement Model 34
3.4 3D Vessel Segmentation 35
3.4.1 Method 36
3.4.1.1 Pre-processing 36
3.4.1.2 Enhanced angiogram synthesis 36
3.4.1.3 Post-processing 38
3.4.2 Experiments 38
3.4.2.1 Competing methods 38
3.4.2.2 Implementation details 38
3.4.3 Results 38
3.4.3.1 Qualitative results 38
3.4.3.2 Quantitative results 39
3.4.4 Discussion 39
3.5 2D Pseudo-modality Generation 42
3.5.1 Method 42
3.5.2 Experiments 43
3.5.3 Results 44
3.5.4 Discussion 44
3.6 Chapter Summary 45

4 Domain Generalization and Domain Adaptation 46

4.1 Introduction 46
4.2 Related Work 49
4.2.1 Domain Generalization 49
4.2.1.1 Data augmentation 49
4.2.1.2 Domain alignment 49
4.2.1.3 Meta learning 50
4.2.2 Unsupervised Domain Adaptation 50
4.2.2.1 Feature alignment 50
4.2.2.2 Image alignment 50
4.3 Explicit shape modelling: Vector field transformer 51
4.3.1 Methods 51
4.3.1.1 Hessian-based vector field 51
4.3.1.2 Vector field transformer 53
4.3.2 Datasets 55
4.3.3 Experiments 56
4.3.3.1 Implementation Details 56
4.3.3.2 Competing Methods 57
4.3.3.3 Source and Target domains 57
4.3.4 Results 58
4.3.4.1 Ablation Study 58
4.3.4.2 Comparison with Competing Methods 59
4.3.5 Discussion 61
4.4 Explicit shape modelling: VesselMorph 65
4.4.1 Method 65
4.4.1.1 Bipolar Tensor Field 65
4.4.1.2 Latent Vessel Representation 67
4.4.1.3 Fusion of Vessel Representations 68

4.4.2 Datasets . . . 69
4.4.3 Experiments . . . 70
4.4.3.1 Compared Methods . . . 70
4.4.3.2 Implementation Details . . . 71
4.4.4 Results . . . 71
4.4.5 Discussion . . . 72
4.5 Implicit shape modelling: MAP . . . 73
4.5.1 Method . . . 74
4.5.1.1 Pseudo-modality Synthesis . . . 74
4.5.1.2 Meta-learning on Anatomy Consistent Image Space . . . 75
4.5.1.3 Structural Correlation Constraints . . . 76
4.5.2 Datasets . . . 77
4.5.3 Experiments . . . 78
4.5.3.1 Implementation Details . . . 78
4.5.3.2 Competing Methods . . . 78
4.5.4 Results . . . 78
4.5.4.1 Ablation Study . . . 78
4.5.4.2 Comparison to Competing Methods . . . 79
4.5.5 Discussion . . . 79
4.6 AdaptDiff . . . 81
4.6.1 Method . . . 82
4.6.2 Datasets . . . 85
4.6.3 Experiments . . . 85
4.6.3.1 Implementation details . . . 85
4.6.3.2 Competing methods . . . 86
4.6.4 Results . . . 86
4.6.4.1 Quantitative results . . . 86
4.6.4.2 Qualitative results . . . 86
4.6.5 Discussion . . . 87
4.7 Chapter summary . . . 89

5 Future work and Conclusion . . . 90

References . . . 93

LIST OF TABLES

Table Page

2.1 Quantitative evaluation 30

3.1 2D FP datasets for pseudo-modality generators training. 34
3.2 Quantitative evaluation table of human and zebrafish segmentation. 41

4.1 Datasets used for vector field transformer 56
4.2 Comparison of model complexisty 57
4.3 Experiment settings for vector field transformer 58
4.4 Dice coefficient of Type I experiment 61
4.5 Accuracy of Type I experiment 61
4.6 Recall of Type I experiment 62
4.7 Hausdorff Distance of Type I experiment 62
4.8 Dice coefficient of Type II experiment 63
4.9 Accuracy of Type II experiment 63
4.10 Hausdorff Distance of Type II experiment 64
4.11 Hausdorff Distance of Type II experiment 64
4.12 Datasets used for VesselMorph 70
4.13 VesselMorph quantitative results 72
4.14 Datasets used in MAP 78
4.15 The ablation study for MAP 79
4.16 The quantitative results for MAP 79
4.17 Datasets used in AdaptDiff 85
4.18 Quantitative results of AdaptDiff 86

LIST OF FIGURES

Figure Page

1.1 Visual examples of color fundus photography . . . 3
1.2 Visual example of FA image . . . 4
1.3 Visual example of OCT and OCT-A . . . 6
1.4 2D depth-projection image for OCT and OCT-A . . . 7

2.1 Examples for in-house human retina OCT dataset . . . 14
2.2 The overall pipeline of self-fusion . . . 16
2.3 Self-fusion for high-noise (HN) single b-scan and low-noise (LN) images . . . 17
2.4 The overall pipeline of PMFN . . . 18
2.5 Fovea denoising results for different input SNR (Excess background trimmed) . . . 20
2.6 Retinal layers in an OCT b-scan . . . 21
2.7 Layer separation analysis . . . 22
2.8 Background (yellow) and foreground (red) ROIs . . . 23
2.9 Quantitative evaluation of denoising results . . . 24
2.10 General workflow for the diffusion probabilistic model. . . . 25
2.11 Fovea denoising results for different input SNR levels and for different t values . . . 28
2.12 Background ENL for different *t* values. . . . 29
2.13 Result comparison with baseline model. . . . 30

3.1 Three manually labelled vessel plexuses . . . 33
3.2 Synthesis model structure . . . 34
3.3 Original *en-face* image and the motion artifact removal result. . . . 36
3.4 Comparison between $\mathbf{x}_i$ and $\mathbf{s}_i$. . . . 37
3.5 Examples of synthesized images. . . . 37
3.6 Human retinal OCT-A segmentation results. . . . 39
3.7 Zebrafish retinal OCT-A segmentation results. . . . 39
3.8 Tractography of the segmentation results to show the vessel connectivity . . . 40
3.9 Quantitative result for human and zebrafish data. . . . 41
3.10 Example of generated pseudo-modalities. . . . 42
3.11 Comparing pseudo-modality with the manual label . . . 43

4.1 Visual examples for domain shift on retinal vessel segmentation . . . 46
4.2 The overall pipeline of an the proposed algorithm . . . 51
4.3 Vector field of different domains . . . 53
4.4 The network architecture of the VFT . . . 54
4.5 Transformer block . . . 55
4.6 T-SNE visualization of datasets used for VFT . . . 56
4.7 Ablation study results for vector field transformer . . . 58
4.8 Qualitative results for vector field transformer . . . 60
4.9 The overall model structure of VesselMorph . . . 65
4.10 The visualization of bipolar tensor field . . . 66
4.11 Qualitative ablation for VesselMorph . . . 70
4.12 Boxplot for ablation results . . . 71
4.13 t-SNE plots of raw data and synthesis data . . . 72
4.14 The key components of MAP . . . 74
4.15 Examples of Dirichlet distribution and corresponding sample images. . . . 76
4.16 Feature clusters and NCC matrix . . . 77
4.17 Domains setting for UDA . . . 82
4.18 Step-wise illustration for the AdaptDiff . . . 82
4.19 Semantic conditional diffusion model . . . 83

4.20 Visual examples for the weakly supervised conditional diffusion 84
4.21 Qualitative results of AdaptDiff 87

CHAPTER 1

Introduction and Specific Aims

As the revolutionary advancement on the computational power have facilitates the direct manipulation on vast amount of data, researchers have explored many algorithms capable of discerning the intrinsic patterns within this data. Such learning-based techniques have profoundly influenced the field of medical image analysis, despite several primary obstacles that hinder their broader application. In my research, I strive to tackle these obstructions to develop deep learning methods that enhance retinal vessel recognition across various image modalities.

1.1 Background

1.1.1 Significance of analysing retinal vessels

Analyzing retinal vessels is crucial in ophthalmology because the vasculature of the retina can reveal significant information about systemic and ocular diseases. It provides valuable information about the vascular health of a patient and can serve as an early indicator of diseases. For example, changes in the retinal vessels are among the earliest signs of diabetic retinopathy (DR) (Cogan et al., 1961; Nguyen and Wong, 2009; Verma et al., 2011), a leading cause of blindness in working-age adults, where high blood sugar levels cause blood vessel damage. Similarly, the hypertensive retinopathy (HR) (Walsh, 1982) can manifest as narrowing, nicking, or swelling of these vessels, mirroring the effects of high blood pressure and offering a glimpse into the patient's cardiovascular health (Wong and McIntosh, 2005). To assess the presence and severity of HR, the ratio between arterio and venous is an important measurement (Irshad and Akram, 2014). Retinal vessel analysis also aids in the diagnosis of vascular occlusions (Mirshahi et al., 2008), such as retinal vein or artery occlusions (Hayreh et al., 2009), which can lead to sudden vision loss. Moreover, age-related macular degeneration (AMD) (Toto et al., 2016), particularly in its 'wet' form, involves abnormal growth of blood vessels under the retina. Retinopathy of prematurity (ROP) (Hellström et al., 2013) is a potentially blinding eye disorder that primarily affects premature infants, particularly those born before 31 weeks of gestation or weighing less than 1250 grams at birth. This condition occurs when abnormal blood vessels grow and spread throughout the retina, the light-sensitive layer of tissue at the back of the eye. These abnormal vessels are fragile and can leak, leading to retinal scarring and, in severe cases, retinal detachment, which can result in blindness. By assessing the structural and functional changes in retinal vessels, clinicians can diagnose, monitor progression, and make treatment decisions for these critical conditions effectively. Therefore, the detailed analysis of retinal vessels is not only fundamental for diagnosing and managing ocular diseases but also offers insights into systemic health, contributing significantly to preventative medicine.

Some research focused on the correlation between retinal vessel abnormalities and cerebrovascular pathology, regarding the alteration of retinal microvascular as a potential biomarker for neural diseases (Moss, 2015; Cabrera DeBuc et al., 2017; Rim et al., 2020). The non-invasive visualization of retinal vessels offers a unique opportunity to predict and understand cerebrovascular conditions such as strokes (Ikram et al., 2006) and dementia (Cheung et al., 2017). For instance, abnormalities like retinal arteriolar narrowing, arteriovenous nicking, and the presence of retinal microaneurysms have been associated with an increased risk of stroke and small vessel disease in the brain. These retinal signs are thought to reflect microvascular damage resulting from hypertension (Keith et al., 1974) and other vascular risk factors that are common to both cerebrovascular and retinal diseases. Moreover, the presence of retinal emboli can indicate atherosclerosis, a major risk factor for stroke. Changes in retinal vessel calibre and the integrity of the blood-retinal barrier may serve as surrogate markers for cerebral small vessel disease, providing insights into the vascular pathologies that underpin both retinal and cerebral conditions. Therefore, by examining the health of retinal vessels, clinicians can potentially identify individuals at higher risk of cerebrovascular events, facilitating earlier interventions and preventative measures to mitigate these risks. This emerging field highlights the importance of integrated care approaches in patients with vascular diseases, emphasizing the retina as a window to the brain's vascular health.

The quantification of these diseases is based on analyzing the morphological features such as the area of foveal avascular zone (FAZ) (Chui et al., 2012), vessel density (You et al., 2017; Lee et al., 2020), vessel diameter index (Tang et al., 2017; Bek, 2017; Toulouie et al., 2022) and vessel tortuosity (Sasongko et al., 2011). All these measurements need to be computed on the binary vessel map or the skeletonized version of it. Nevertheless, the manual labeling can be extremely laborious and time-consuming. Unlike the brain tissue and abdominal organs, the retinal vessels are usually ubiquitous in the image and many small capillaries are particularly hard to discern. Figure 1.2 shows an example of labeled 2D high-resolution fluorescein angiography. Ding et al. (2020a) annotated all the small capillaries in the image which takes significant endeavor. Note that there are only eight images in this dataset. The 3D scenario can be even worse, and thus there is no public 3D dataset with manual labels. Therefore, an automated vessel segmentation algorithm is required since the manual segmentation is not feasible for the complex retinal vasculature.

1.1.2 Deep learning approach for medical image analysis

Since Krizhevsky et al. (2012) presented a convolution neural network (CNN) dubbed *AlexNet* in late 2012 and won the ImageNet challenge by a large margin, the deep learning techniques have permeated the entire field of both computer vision and medical image analysis for the last decade (Litjens et al., 2017). Al-

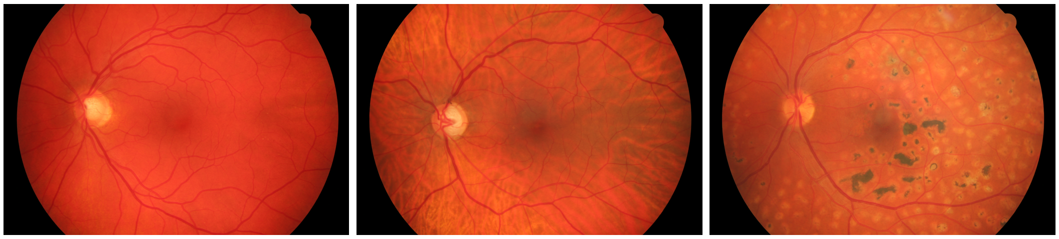

(a) healthy (b) glaucoma (c) diabetic retinopathy

Figure 1.1: An example of healthy retina compared to retinas with glaucoma and diabetic retinopathy under fundus photography. The images are from the high resolution fundus (HRF) dataset (Budai et al., 2013).

though the traditional model-based algorithms are rigorous in their derivation, they often rely on simplistic assumptions about the image characteristics, which may not hold true for complex medical images, leading to reduced accuracy and robustness. In contrast, deep learning methods offer significant advantages, primarily due to their ability to directly learn from a large amount of data, capturing intricate patterns. This process only depends on the training data instead of an oversimplified assumption, and the large amount of learnable parameters allow the neural networks to model a highly non-linear mapping function between the input and the desired output. The performance on tasks including segmentation (Hesamian et al., 2019), object detection (Kaur et al., 2021) and classification (Yadav and Jadhav, 2019) can be greatly improved by training a simple neural network.

Another watershed moment for the development of the learning-based algorithm was the introduction of the generative adversarial network (GAN) (Goodfellow et al., 2014). Deep models are able to delineate the statistical distribution of high dimensional data and sampling on this domain enables the generation of unseen images. GAN has significant impact on medical image analysis as it enables cross-modality image synthesis. Moreover, other tasks like denoising (Sagheer and George, 2020), super-resolution (Yang et al., 2023) and harmonization (Dewey et al., 2019) also greatly benefit from this innovation.

In my research, I investigate several deep learning applications on image denoising, vessel segmentation and image synthesis to advance the retinal vessel cognition on the commonly used ophthalmic imaging modalities including fundus photography (FP), fluorescein angiography (FA), optical coherence tomography (OCT) and OCT angiography (OCT-A).

1.1.3 Major ophthalmic imaging modalities

1.1.3.1 Fundus photography

Color fundus photography (FP) is a specialized type of medical imaging that captures a sharp view of the retina, the retinal vasculature, and the optic nerve head (optic disc) from which the retinal vessels enter the eye. This technique utilizes a fundus camera—a complex optical system consisting of a low-power micro-

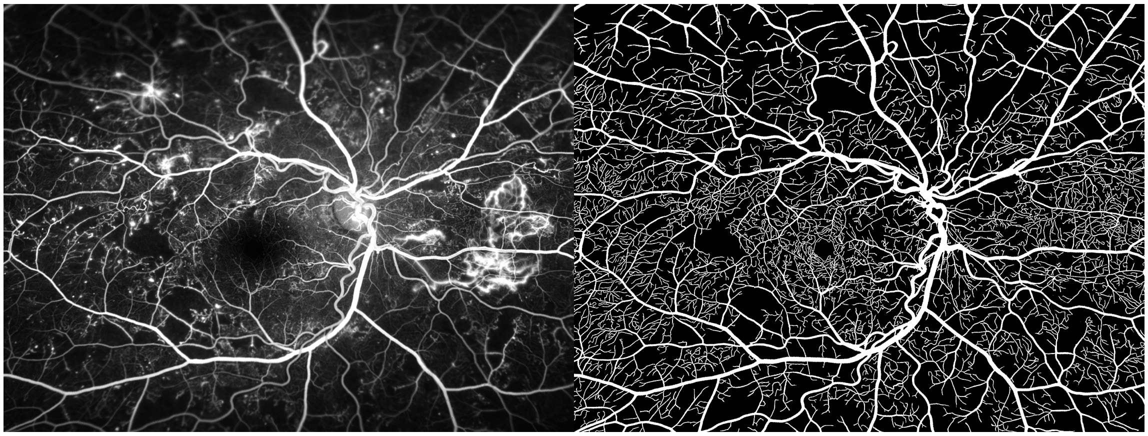


Figure 1.2: Visual example of fluorescein angiography. The image is from the RECOVERY-FA19 dataset (Ding et al., 2020a). **Left**: FA image, **Right**: the manual label for retinal vessels.

scope with an attached camera—designed to focus light through the pupil and capture clear images of the fundus. The camera emits a flash of light, which reflects off the retina and is then captured by the camera lens, creating a photograph of the back of the eye. This imaging is crucial in diagnosing and monitoring various ocular diseases such as diabetic retinopathy (Liesenfeld et al., 2000), glaucoma (Bock et al., 2007), age-related macular degeneration (Midena et al., 2020), and retinal detachment. Figure 1.1 is a visual example of healthy retina and pathological phenotypes in which features like dot hemorrhages, exudates and morphological change in optic disc and microvasculature are usually spotted. FP is a non-invasive method that provides high-resolution 2D images that help in the detailed examination of the retinal structure, facilitating the identification of abnormalities and diseases at early stages. It serves as a permanent record of the patient's retinal condition and allows for the tracking of changes over time, enabling ophthalmologists to monitor the progression of ocular diseases and evaluate the effectiveness of treatments.

Because the FP images are easy to get access to and they primarily visualize the large retinal vessels that are tractable for manual annotation, there are many public annotated FP datasets (Staal et al., 2004; Hoover et al., 2000; Li et al., 2020b) providing the segmentation maps of retinal vessels and optic discs. The sufficient amount of labeled data enables more research to be conducted on deep learning approaches for the vessel segmentation task. However, the rich features contained in FP can potentially occlude the vasculature and the image contrast is usually far from optimal to visualize the vessels. Moreover, FP does not include any depth information and the small capillaries in deeper layers (e.g., outer plexiform layer) of the retina are not visible.

1.1.3.2 Fluorescein angiography

Fluorescein angiography (FA) is a diagnostic procedure used in ophthalmology to visualize the blood vessels of the retina and choroid, the layers of the back of the eye. This technique involves the intravenous injection of fluorescein, a fluorescent dye, which travels through the bloodstream and into the blood vessels of the eye. As the dye circulates, a specialized camera equipped with filters that highlight the fluorescence takes a series of rapid-sequence photographs. These images capture the flow of blood and reveal any blockages, leaks, or abnormalities in the vasculature. FA is the gold standard for in vivo evaluation of the retinal circulation (Spaide et al., 2015) and provide significantly better 2D visualization of the retinal vasculature than FP. Figure 1.2 shows an example of FA image. Obviously, the FA is exceptionally good at visualizing vessels in high resolution and with decent image contrast. The fluorescent dye highlights the retinal and choroidal vascular abnormalities such as diabetic retinopathy, macular degeneration, and retinal vein occlusions. This dye-based angiography has been regarded as the gold standard modality for evaluating retinal and choroidal vascular pathologies.

Despite its widespread success, FA is invasive and time-consuming, in addition to having the potential for allergic reactions to the dyes (Yannuzzi et al., 1986). Similar with the fundus photography, FA is only a two-dimensional image focusing on the superficial retinal circulation, without the ability to visualize the deeper capillary structures. Any leakage of the dye can also occlude the vessels in certain regions. These limitations spurred the development of faster, safer imaging techniques, capable of effectively imaging both the retinal and choroidal circulation in 3D.

1.1.3.3 Optical coherence tomography

Optical coherence tomography (OCT) is a non-invasive imaging technique widely used in ophthalmology to capture high-resolution cross-sectional images of the retina, optic nerve, and anterior segment of the eye. OCT works on the principle of low-coherence interferometry (Huang et al., 1991). It utilizes near-infrared light to create detailed images of the eye's internal structures by measuring the echo time delay and intensity of the light as it reflects off different tissue layers. Compared with the aforementioned imaging techniques FP and FA, OCT can provide a volumetric representation of the retina with rather high speed. It takes only a few minutes to scan one eye. In the top row of Figure 1.3, I show an example of the OCT volume of human retina. The red dashed line indicates a cross-sectional image called b-scan. The b-scan image clearly shows the layer structure of the retina which is particularly effective in detecting abnormalities within the retinal layers that may indicate the presence of diseases such as diabetic retinopathy (Virgili et al., 2015), AMD (Keane et al., 2012) and retinal detachment (Abouzeid and Wolfensberger, 2006). By assessing the thickness and morphology of the retinal layers, OCT helps in diagnosing the specific type of retinal pathology

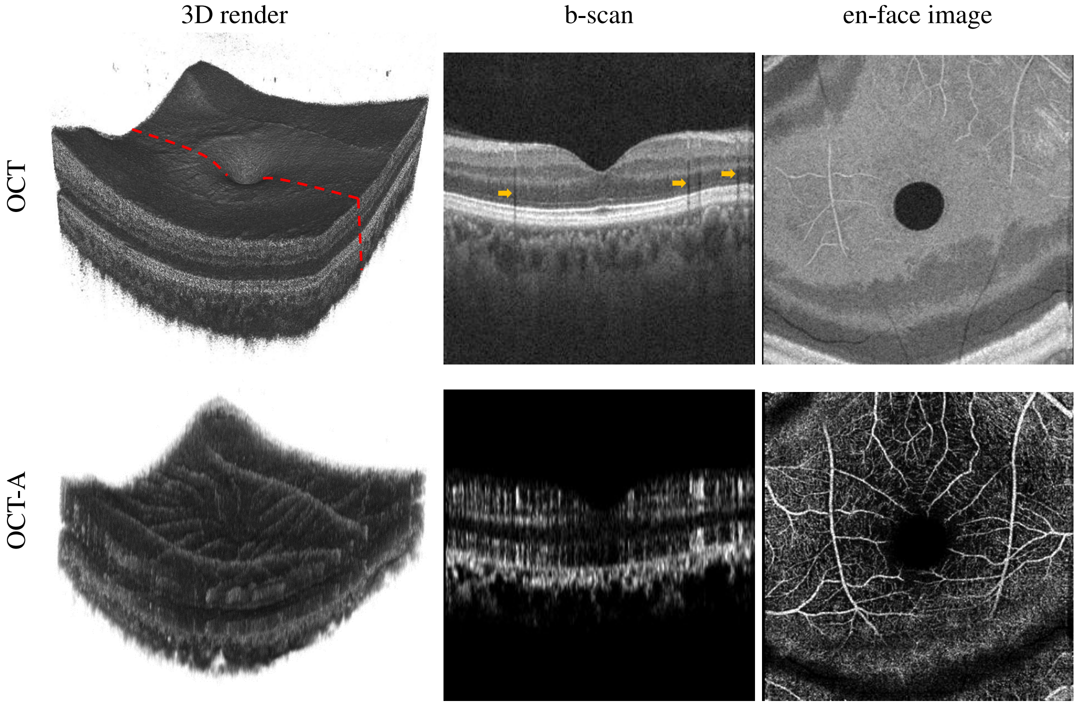
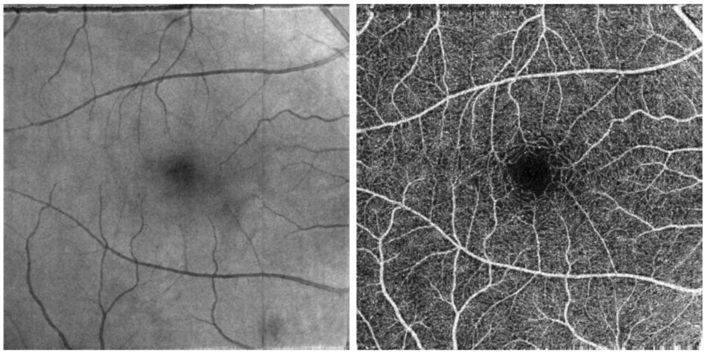


Figure 1.3: Visual example of OCT and OCT-A. The red dash line indicates a b-scan. The yellow arrows in the OCT b-scan image mark the shadow beneath large vessels.
The en-face image is a 2D image on the depth axis. The OCT shows the tissue layers within the retina, whereas the OCT-A highlights the vascular structure. These volumes are from the OCTA-500 dataset (Li et al., 2020b).

and monitoring the efficacy of ongoing treatments. Note that in the b-scan, the intersection with larger vessels in shallow layers (e.g., inner plexiform layer) is visible as they have shadow beneath them (the thin vertical dark strips highlighted with the yellow arrows). The 2D image along the depth axis refers to the en-face image which is illustrated in the right column. In many applications, it is common to compute the depth-projection to visualize the retinal vasculature in a 2D image (Figure 1.4). The vessels in the depth-projection of OCT volume have low intensity since their dark shadow blocks the bright layers underneath.

The OCT angiography (OCT-A) is a modality derived from OCT that is specifically designed to visualize the vascular structures. OCT-A leverages the motion contrast of flowing blood cells compared to static tissue to capture detailed images of blood flow. Practically, this process involves taking multiple OCT b-scans at the same spatial location over a short period. By computing the variance of these sequential b-scans, the vessels are decoupled from the surrounding stationary tissue. By its nature, the OCT-A is sensitive to unresolved motion artifacts and speckle noise which can also induce high variance across repeated b-scans. The bottom row of Figure 1.3 is the corresponding OCT-A of the OCT volume in the top row. Clearly, the vessels stand out from the layer tissue in the OCT-A image. In the depth-projection 2D image (Figure 1.4), there are more



(a) depth-projection of OCT (b) depth-projection of OCT-A.

Figure 1.4: 2D depth-projection image for OCT and OCT-A

smaller capillaries visible than in the OCT counterpart. Therefore, this groundbreaking imaging technology allows clinicians to observe blood flow in the retinal vessels with great precision, highlighting abnormalities that might indicate various retinal diseases. For instance, OCTA is crucial in detecting areas of non-perfusion, abnormal vessel growth, and microaneurysms associated with diabetic retinopathy (Schreur et al., 2019). It is also instrumental in diagnosing and monitoring choroidal neovascularization (Chen et al., 2016). Furthermore, OCT-A can help detect changes in vessel density and diameter that accompany conditions like retinal vein occlusion or glaucoma (Rao et al., 2020).

1.1.4 Challenges

To develop a deep learning approach for automatic retinal vessel segmentation on the aforementioned image modalities, there are three major obstacles that should be addressed. These challenges are of the primary concerns for almost all learning-based medical image analysis methods, so the solutions to tackle these issues will not only benefit our specific project or just retinal imaging, but also have a wide-reaching impact on the overall medical image analysis field.

1.1.4.1 Low image quality

Unlike normal cameras used for natural images, visualizing cross-sectional images of tissues in a non-invasive manner requires a sophisticated imaging technique that can potentially have a trade-off in terms of imaging time and image quality. As an example, the OCT is based upon the low-coherence interferometry which gives rise to speckle noise that can significantly degrade the image quality. This granular noise is primarily caused by the coherent nature of the laser light used in OCT imaging. OCT uses coherent light (light with waves that are phase-aligned) to illuminate the tissue being imaged. As this coherent light penetrates the tissue, it scatters in various directions due to interactions with different tissue components at microscopic scales. Then the scattered light waves reflect back to the OCT sensor, where they interfere with each other. Some of these light waves constructively interfere and others destructively interfere. This interference forms the speckle pattern detected by the OCT system and becomes part of the reconstructed image. Furthermore, handheld OCT devices (Malone et al., 2019), which can induce more speckle noise and motion artifacts.

In clinical practice, the thickness of the retina layers and the vascular system are important for observing ocular disease. The speckle noise in single frame b-scans makes the border of layers unclear so that it is hard to distinguish adjacent layers. The noise also produces bright dots and dark holes that can hurt the homogeneity of layers and affect the visibility of the small vessels within them. Consequently, it is hard to conduct image analysis on the high-noise OCT images for either direct diagnosis or other tasks like vessel/layer segmentation.

1.1.4.2 Insufficient manual annotations

One of the significant challenges in utilizing deep learning methods for medical image segmentation tasks is the scarcity of manually annotated datasets. These annotations are crucial as they provide the ‘ground truth’ used to train models to accurately interpret and segment medical images. Manual annotations require extensive time and effort from highly skilled professionals, such as radiologists or pathologists, who must delineate structures precisely in each image. This process is not only labor-intensive but can also be subject to variability between raters. Unlike other segmentation tasks on brain tissue or abdomen organs, the retinal vessels are ubiquitous in the 3D volume/2D image which makes them even harder to label. Moreover, due to the limitations of image quality, many capillaries are too subtle to discern, which induces severe disagreement among different raters. As a result, the limited availability of annotated datasets can hinder the development of robust deep learning models, as these models typically require large amounts of data to achieve high accuracy and generalizability. Addressing this bottleneck is critical for advancing the capabilities of deep learning in medical image analysis, prompting a need for unsupervised or semi-supervised methods that can reduce the requirement for manual labels.

1.1.4.3 Domain shift

Domain shift represents a formidable challenge in deploying deep learning models for medical image segmentation across different clinical settings. This issue arises when a model, trained on a specific dataset obtained from one set of equipment or demographic, performs poorly when applied to data from different center, equipment, or patient populations. Such shifts can occur due to variations in imaging protocols, scanner hardware, or even subtle differences in patient characteristics and disease presentations. These discrepancies can lead to significant degradation in model performance because the features learned by the model may not be universally applicable across different domains. For instance, a model trained on OCT images from one type of scanner might struggle with images from a newer model that uses different light wavelengths or detection technologies. Addressing domain shift is critical for ensuring that deep learning models are robust and reliable across various clinical environments, necessitating techniques like domain generalization and domain adaptation to bridge the gap between different datasets and enhance model generalizability.

1.2 Specific Aims

1.2.1 OCT image denoising

To address the high noise level of OCT images, acquiring multiple b-scan frames at the same spatial location and averaging these repeated frames is the mainstream technique for OCT denoising. Theoretically, the more repeated frames are acquired, the closer their mean can be to the ideal ground truth. However, this increases the imaging time linearly, and can cause discomfort to patients as well as increase motion artifacts. Other hardware-based OCT denoising methods including spatial (Avanaki et al., 2013) and angular averaging (Schmitt, 1997) will similarly prolong the acquisition process. Ideally, an image post-processing algorithm that applies to a single frame b-scan is preferable. Therefore, I develop two methods for OCT b-scan denoising in Chapter 2.

1.2.2 Retinal vessel enhancement

To tackle the lack of manual annotations on retinal vessels, I leverage deep neural networks to generate vessel enhanced angiograms which provide improved visibility of retinal vessels and suppressed speckle noise and background artifacts. The algorithm is desired to be applicable on different modalities such as fundus photography and OCT-A. With the synthetic angiogram, it is possible to obtain good binary vessel segmentation maps via traditional model-based algorithms. Although the method is designed to work on 2D images, I extend the application to unsupervised volumetric vessel segmentation on 3D OCT-A data. This work is presented in Chapter 3.

1.2.3 Domain adaptation and generalization

To deal with the domain shift between different datasets, I propose both domain generalization and domain adaptation methods in Chapter 4. For domain generalization, I focus on modeling the domain-invariant feature, which is the morphological structure of the vasculature. I observe that this can be achieved in either explicit or implicit delineation. As for the domain adaptation, I propose to train the model on synthesized images, which can significantly reduce the need for labeled data on any unseen target domain.

1.3 Overview of the thesis

In Chapter 2, 3, 4, I detail the methodologies related to the three specific aims of my study. Each chapter begins with an overview of the task's general background, followed by a discussion of key prior works that address the specific problem. Subsequently, I describe the datasets utilized in my experiments. Then I present the novel methods developed in my work. Following that, both the qualitative and quantitative results of these methods are shown and discussed. Each chapter concludes with a summary that encapsulates the work done with regard to each aim. In Chapter 5, I explore the potential future research directions based on my study. Finally, I talk about the major takeaways and overarching gains from my Ph.D research.

CHAPTER 2

OCT Denoising

2.1 Introduction

Optical coherence tomography (OCT) (Huang et al., 1991; Fercher et al., 2003) is a non-invasive imaging technique based upon low-coherence interferometry. By providing high resolution cross-sectional visualization of retinal tissue in-vivo, OCT has revolutionized the clinical practice of ophthalmology (Adhi and Duker, 2013). Depth-resolved retinal images are able to provide essential biomarkers such as morphological change of retinal layers that are used for diagnosis and monitoring of ocular pathologies including glaucoma (Bowd et al., 2000; Medeiros et al., 2005), age-related macular degeneration (AMD) Farsiu et al. (2014); Srinivasan et al. (2006) and diabetic retinopathy (DR) (Chiu et al., 2015). However, due to limited spatial-frequency bandwidth, detection of coherent light waves has an inherent characteristic of speckle (Schmitt et al., 1999; Wong et al., 2010). Although the speckle pattern contains some texture information of the biological tissue, it can severely degrade the image quality by occluding essential anatomical structures such as retinal vessels and thin layers. Thus, a proper denoising method is paramount for ophthalmic application on OCT.

Averaging multiple repeated 2D image acquisitions (b-scans) taken at the same spatial location is the mainstream technique for OCT noise reduction (Sander et al., 2005; Sakamoto et al., 2008). The more repetitions are acquired, the closer their mean can be to the ideal ground truth. However, this approach prolongs the imaging time linearly, and can cause discomfort to patients as well as increase motion artifacts. Moreover, the registration error can potentially cause a blurring effect that is detrimental to small features. Therefore, an image post-processing method is preferred to denoise the OCT b-scans.

Deep learning has become the state-of-the-art in many image processing tasks and shown great potential for image noise reduction. Many learning-based approaches have been investigated for the OCT denoising (Ma et al., 2018; Devalla et al., 2019; Mao et al., 2019; Fan et al., 2020). In the following sections, I first discuss literature of the learning-based denoising algorithms categorized with regard to different types of supervision. Then I introduce a non-learning-based self-fusion algorithm (Oguz et al., 2020) for OCT b-scan denoising. Based upon self-fusion, I propose a supervised learning method and an unsupervised learning method for OCT b-scan denoising in Sec. 2.6 and Sec. 2.7 respectively.

2.2 Related Work

In this section, I discuss recent approaches of deep learning methods for OCT denoising. These methods are categorized by the different settings in supervision settings.

2.2.1 Supervised methods

2.2.1.1 Direct prediction

Supervised learning often requires a ground truth $\mathbf{y}$ for each input data $\mathbf{x}$. However, in medical image noise reduction problem, the noise-free ground truth is not accessible. The most commonly used substitution of the ideal image is the average of multiple b-scan acquisitions at the same position. It is a hardware-based method that can always provide a robust result with high signal-to-noise ratio (SNR). It is widely used as ground truth in supervised deep learning OCT denoising e.g., (Devalla et al., 2019; Gour and Khanna, 2019; Qiu et al., 2020; Yu et al., 2018; Chen et al., 2020c). With these methods, the model is trained to directly predict the low-noise image given a noisy input and thus they are classified as direct prediction.

Some commercial OCT imaging devices may not be able to give repetitions of b-scans in a volume. Ma et al. (2018) present an alternative way that can work for any type of scanner. Instead of repeating b-scans, M volumes are obtained from the same eye. One of the volumes V^t is selected as the template. They notate the i^{th} b-scan in the template volume as $\mathbf{x}_i^t$. For each of the rest $M-1$ volumes V^k where $k \in [1, M-1]$, they take $\mathbf{x}_i^k$ and its $2N$ neighboring b-scans $\mathbf{x}_{i-N}^k, \dots, \mathbf{x}_{i+N}^k$ as a set of candidates. Then they register all the $2N+1$ candidate b-scans to the template $\mathbf{x}_i^t$. Among all the registered candidates, they take the average over L images with the highest mean structural similarity index (MSSIM) (Wang et al., 2004) to get the low noise ground truth. For the notation here, the superscript indicates the volume index while the subscript is the b-scan index.

2.2.1.2 Residual prediction

Speckle is considered to be a kind of multiplicative noise. Limited by the dynamic range of display, OCT image intensity is the logarithmic of the original signal. In this way, the multiplicative speckle is converted to additive noise which can be expressed as $\mathbf{x} = \mathbf{y} + \varepsilon$, where $\mathbf{y}$ is the noise-free signal, ε is the speckle noise and $\mathbf{x}$ is the noisy image. Given the input image $\mathbf{x}$ and its corresponding low-noise image $\mathbf{y}$, the noise can be regarded as the residual $\varepsilon = \mathbf{y} - \mathbf{x}$. Other than the conventional way of paring data $(\mathbf{x}, \mathbf{y})$, an alternative is to learn the noise from the input $(\mathbf{x}, \varepsilon)$. Then the deep learning model is trained to predict the noise ε given noisy image $\mathbf{x}$. This residual prediction method is proposed in the works (Wei et al., 2018; Shi et al., 2019; Cai et al., 2018; Chen et al., 2020b; Xu et al., 2020a).

2.2.2 Unsupervised methods

2.2.2.1 Unpaired data

As stated before, acquiring multiple images to create the low-noise ground truth for a specific input image is resource intensive. But it may be relatively easier to get a set of high-noise data $\mathscr{D}^h = \{\mathbf{x}_1^h, \dots, \mathbf{x}_N^h\}$ and a set of low-noise images $\mathscr{D}_l = \{\mathbf{x}_1^l, \dots, \mathbf{x}_M^l\}$ without one-to-one correspondence. A similar problem exists prevalently in machine translation as there is no strict word-to-word matching between different languages. Inspired by the solution of learning from unpaired data in machine translation area (Artetxe et al., 2017; Lample et al., 2017), Adiga and Sivaswamy (2018) present a shared encoder (SE) architecture that can be trained by unpaired high-noise and low-noise data. Basically, one shared encoder $E(\cdot)$ is implemented to map both the high-noise images $\mathbf{x}^h$ and the low-noise images $\mathbf{x}^l$ to the same latent space. Then two decoders $D^h(\cdot)$ and $D^l(\cdot)$ are used to reconstruct the noisy and clean output, respectively. For a low-noise input $\mathbf{x}_i^l$, its corresponding noisy version can be generated by $\hat{\mathbf{x}}^h = D^h\left[E(\mathbf{x}_i^l)\right]$. Then $\{\hat{\mathbf{x}}_i^h, \mathbf{x}_i^l\}$ is regarded as a pseudo-pair of training data for $D^l\left[E(\cdot)\right]$. By iteratively training, the converged autoencoder $D^l\left[E(\cdot)\right]$ is the denoising model acquired. Huang et al. (2020); Guo et al. (2019) leverage the idea of shared encoder in a CycleGAN (Zhu et al., 2017).

2.2.2.2 Noise2Noise

Lehtinen et al. (2018) present a Noise2Noise (N2N) image restoration in which they propose to train an image denoising model using a pair of noisy data $(\mathbf{x}_h^1, \mathbf{x}_h^2)$ as long as they share the same underlying signal structure. In the work Wu et al. (2019), Wu et al. mathematically prove that the parameters in N2N model will converge to that of a conventional noise-to-clean model under the assumption that the noise is independent and zero-mean. This conclusion has great impact on the research of medical image restoration considering the cost of high-quality image acquisition. Mao et al. (2019) use two repeated b-scans as input and target to train a denoising network. Although the speckle noise is usually approximated by Gamma or Rayleigh distribution that are not zero-mean, the N2N approach can still, empirically, achieve decent performance in the speckle removal task as stated in the study (Qiu et al., 2021; Gisbert et al., 2020; Huang et al., 2021).

2.3 Datasets

In the OCT denoising project, I use the in-house data acquired at the DIIGI Lab at Vanderbilt University obtained from the fovea and the optic nerve head (ONH) of a single human retina (Malone et al., 2019). For each region, two volumes are acquired at three different noise levels (SNR=92dB, 96dB, 101dB). Each raw OCT volume is in shape $[N,H,W] = [500,1024,500]$ where N is the number of b-scans; H and W represent the height and width of each 2D b-scan. For each b-scan, there are 5 repeated frames taken at the same position so that a 5-frame-average can be utilized as the low-noise ground truth to evaluate denoising performance. The Figure 2.1 illustrates the high-noise (HN) b-scans in different SNR and their corresponding low-noise (LN) images. Since all these volumes are acquired from a single object, to avoid information leakage, I denoise the fovea volumes by training on ONH data, and vice versa.

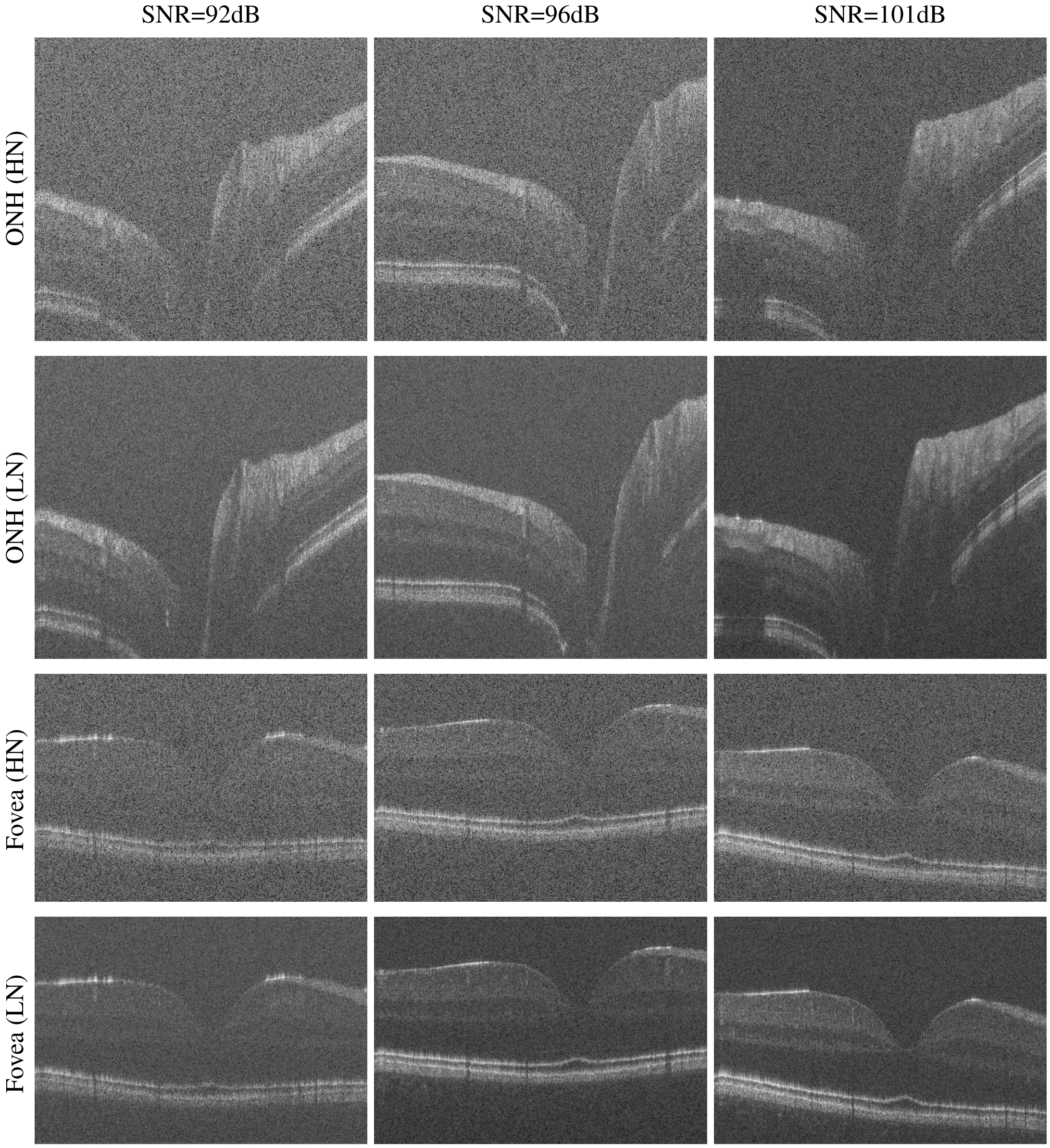


Figure 2.1: Examples for in-house human retina OCT dataset

2.4 Evaluation metrics

In this section, I introduce several metrics that are commonly used to validate the denoised image quality. Note that many of them are not strictly defined since the testing condition can vary in different tasks. For example, in my experiment, there is no noise-free reference image available, and thus the definition of signal-to-noise ratio needs to be adjusted accordingly. I assume that the foreground of the reconstructed image contains pure signal while the background has pure noise. In this way, the SNR can be approximated by taking regions of interest from foreground and background respectively. A straightforward way to define

SNR (Devalla et al., 2019) is:

$$SNR = 10\log_{10}\left[\frac{\sum_i^H \sum_j^W [\mathbf{f}(i,j)]^2}{\sum_i^H \sum_j^W [\mathbf{b}(i,j)]^2}\right] \tag{2.1}$$

where $\mathbf{f}(i,j)$ is the pixel intensity in foreground window and $\mathbf{b}(i,j)$ is background pixel intensity. Another global metric PSNR is defined by:

$$PSNR = 10\log_{10}\left[\frac{HW max[\mathbf{f}(i,j)]^2}{\sum_i^H \sum_j^W [\mathbf{b}(i,j)]^2}\right] \tag{2.2}$$

where the numerator is the maximum intensity of the image, the denominator is the mean square error with regard to the reference background ROI.

Contrast-to-noise ratio (CNR) Bao and Zhang (2003) indicate the contrast between the signal region and the background region. It also requires taking ROI from the foreground and background. Usually it is defined as Ma et al. (2018):

$$CNR = 10\log_{10}\left(\frac{|\mu_f - \mu_b|}{\sqrt{\sigma_f^2 + \sigma_b^2}}\right) \tag{2.3}$$

where μ_f and σ_f are the mean and variance of the foreground ROI while μ_b and σ_b are those of the background region. In some papers Gour and Khanna (2019); Devalla et al. (2019); Yu et al. (2018); Fan et al. (2020); Wei et al. (2018), the authors use $\sqrt{0.5(\sigma_f^2 + \sigma_b^2)}$ as the denominator. To make the index less sensitive to bias, some authors Guo et al. (2020); Huang et al. (2019) take multiple foreground regions and compute the mean CNR.

The SSIM Wang et al. (2004) is used to measure the structure similarity between the denoised image and the reference.

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{2.4}$$

where x, y refer to the denoised image and the reference image, μ, σ are mean and variance, c_1, c_2 are constant to numerically stabilize the division.

To numerically measure the smoothness of the homogeneous signal region, the mean to standard deviation ratio (MSR) Cincotti et al. (2001) and equivalent number of looks (ENL) are applied.

$$ENL = (MSR)^2 = \left(\frac{\mu_f}{\sigma_f}\right)^2 \tag{2.5}$$

2.5 Preliminary: Self-fusion

In this section, I first introduce *self-fusion* which is a non-learning-based OCT denoising method proposed by Oguz et al. (2020). Self-fusion is inspired by a well-known multi-atlas label fusion method for segmenta-

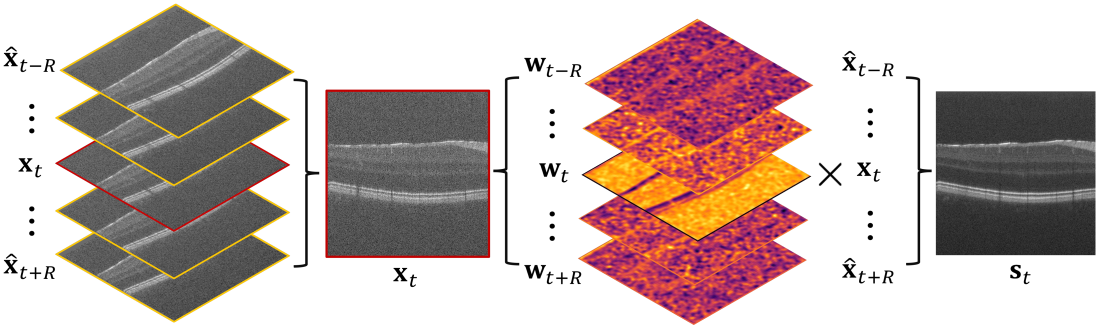


Figure 2.2: The overall pipeline of self-fusion

tion named joint label fusion (JLF) (Wang et al., 2012). In JLF, a library of K atlases with known segmentation maps $\{\mathbf{x}_k, \mathbf{y}_k\}|_{k=1}^{K}$ are registered with a target image $\mathbf{x}_{target}$. The weighted sum of the deformed segmentation maps $\{\hat{\mathbf{y}}_k\}|_{k=1}^{K}$ provides the consensus segmentation $\mathbf{y}_{target}$ for the target image.

$$p(\mathbf{y}_{target}(i,j) = l) = \sum_{k=1}^{K} \mathbf{w}_k(i,j) p(\hat{\mathbf{y}}_k(i,j) = l), \tag{2.6}$$

where l represents the multi-class label and $l \in \{1,\ldots,L\}$ and $\mathbf{w}_k$ is the weight map for the k^{th} atlas. $p(\hat{\mathbf{y}}_k(i,j) = l)$ represents the probability for pixel (i,j) to be labeled as l. To account for the possibility that multiple atlases introduce similar segmentation errors, a $K \times K$ matrix $\mathbf{M}_{ij}$ is introduced to indicate the correlation between atlases. If k_1 and k_2 are two atlases in the library, then

$$\mathbf{M}_{ij}(k_1,k_2) = \alpha \mathbf{M}_0 + \left[|\tilde{\mathbf{x}}_{k_1}(i,j) - \tilde{\mathbf{x}}_{target}(i,j)| \cdot |\tilde{\mathbf{x}}_{k_2}(i,j) - \tilde{\mathbf{x}}_{target}(i,j)| \right]^{\beta} \tag{2.7}$$

where $\alpha \mathbf{M}_0$ is a term introduced to prevent an ill-posed matrix, such that α is a small positive value and $\mathbf{M}_0$ is a $K \times K$ matrix with every entry equal to 1. $\tilde{\mathbf{x}}_k(i,j)$ and $\tilde{\mathbf{x}}_{target}(i,j)$ represent a small patch centered at pixel (i,j). β is a system parameter. To minimize the expectation error, the weight maps $\mathbf{w}$ can be computed by

$$\mathbf{w}(i,j) = \frac{\mathbf{M}_{ij}^{-1}\mathbf{1}}{\mathbf{1}^t \mathbf{M}_{ij}^{-1}\mathbf{1}} \tag{2.8}$$

where $\mathbf{1}$ and $\mathbf{w}(i,j)$ are $K \times 1$ vectors; all entries of $\mathbf{1}$ are equal to 1.

Self-fusion extends the JLF to image synthesis that does not require atlas segmentation maps. The overall workflow is shown in Figure 2.2. Instead of using an external group of atlases, self-fusion regards adjacent b-scans in the OCT volume as a group of atlases. This is based on the assumption that the major anatomical structures (e.g., retinal layers) are not changing abruptly in consecutive b-scans. Therefore, the general structure in all atlases are similar. The number of atlases is defined by the radius R of the neighborhood.

In other words, for the target 2D b-scan $\mathbf{x}_t$ (t is the index of the b-scan in the OCT volume), the atlases are $\{\mathbf{x}_{t-R},\ldots,\mathbf{x}_{t+R}\}$ including $\mathbf{x}_t$ itself. Similar with the JLF, all the neighboring b-scans are registered with $\mathbf{x}_t$, and a self-fused patch centered at pixel (i,j) can be is acquired by weighted sum

$$\mathbf{s}_r(i,j) = \sum_{r=-R}^{R} \mathbf{w}_r(i,j)\hat{\mathbf{x}}_r(i,j) \tag{2.9}$$

Here, the $\mathbf{W}_r$ represent the weight map corresponding to each deformably registered atlas $\hat{\mathbf{x}}_r$. In my experiment, I set $R = 3$. In Figure 2.3 self-fusion result for both high-noise (HN) and low-noise (LN) OCT volume. Note that the low-noise image is the mean of 5 repeated b-scan frames.

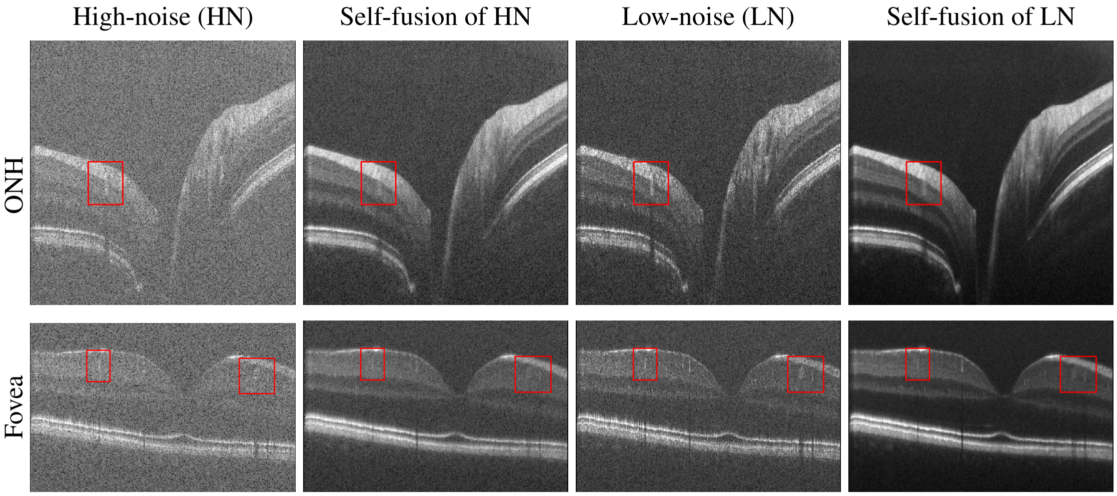


Figure 2.3: Self-fusion for high-noise (HN) single b-scan and low-noise (LN) images

The self-fusion results illustrate that it can effectively suppress the noise in the background and significantly improve the separation between different layers since the retinal layers are relatively consistent in a small adjacency. However, it can potentially over smooth some small features (e.g., vessels) that only exist in a few slides within the neighborhood. The red boxes in Figure 2.3 highlight some examples where the retinal vessels are weakened in terms of contrast after self-fusion. Moreover, computing patch-based similarity between the target b-scan with multiple adjacent frames requires longer processing time. Therefore, I develop two different learning-based algorithms to further improve the OCT denoising performance based upon the self-fusion method.

2.6 Supervised-learning Method: PMFN

In Sec. 2.5, a non-learning-based self-fusion method is introduced for OCT b-scan denoising. Comparing with the low-noise (LN) image which is commonly used as the ground truth for learning-based denoising algorithms, the self-fusion results have better noise suppression and layer separation. However, the LN image is more precise in smaller features such as retinal vessels. In this section, I present a deep learning approach dubbed pseudo-multimodal fusion network (PMFN) that takes the advantages from both LN and self-fusion of LN to reduce the speckle noise in OCT b-scans. In this study, the self-fusion of LN b-scans is regarded as a pseudo-modality denoted as $\mathbf{s}_i$ where the subscript i is the index of the b-scan in the OCT volume. The HN and LN images are denoted by $\mathbf{x}_i$ and $\mathbf{y}_i$ respectively. The overall pipeline is illustrated in Figure 2.4. Specifically I train two networks: Network I for pseudo-modality creation and Network II for pseudo-multimodal fusion.

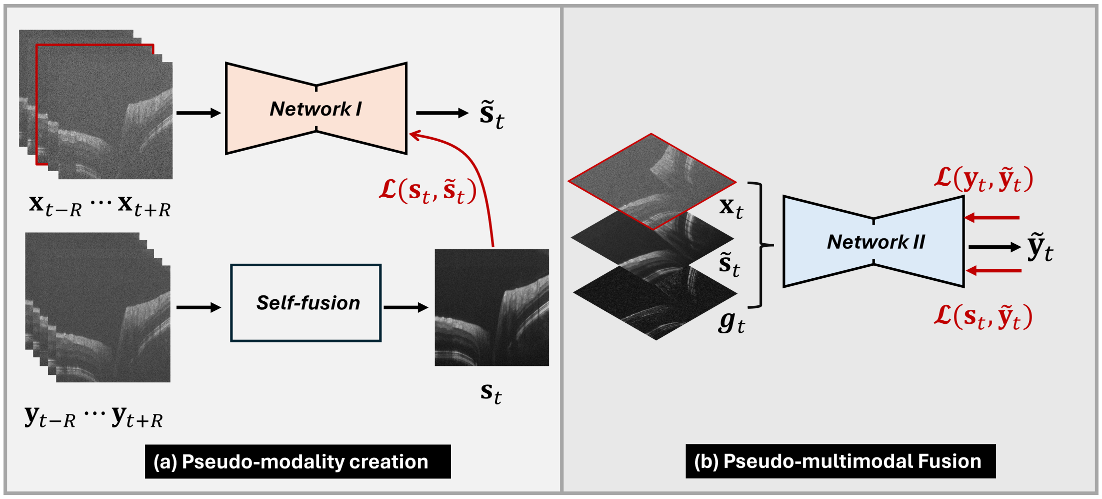

Network I
$\tilde{\mathbf{s}}_t$
$\mathbf{x}_{t-R} \cdots \mathbf{x}_{t+R}$
$\mathcal{L}(\mathbf{s}_t, \tilde{\mathbf{s}}_t)$
Self-fusion
$\mathbf{s}_t$
$\mathbf{y}_{t-R} \cdots \mathbf{y}_{t+R}$
(a) Pseudo-modality creation
$\mathbf{x}_t$
$\tilde{\mathbf{s}}_t$
g_t
Network II
$\mathcal{L}(\mathbf{y}_t, \tilde{\mathbf{y}}_t)$
$\tilde{\mathbf{y}}_t$
$\mathcal{L}(\mathbf{s}_t, \tilde{\mathbf{y}}_t)$
(b) Pseudo-multimodal Fusion

Figure 2.4: The overall pipeline of PMFN

2.6.1 Method

2.6.1.1 Network I: Pseudo-modality creation

To reduce the processing time, I propose to train a neural network (Network I) to replace the self-fusion. Run time for generating a self-fusion image of a b-scan drops from 7.303 ± 0.322s to 0.253 ± 0.005s. The idea allows us to also improve the quality of my pseudo-modality, by using $\mathbf{s}_t$, the self-fusion of LN $\mathbf{y}_t$ images, rather than that of HN images. Thus, Network I maps a stack of consecutive HN b-scans to self-fusion of LN. Note that the number of the input b-scans is $2R+1$ where R is the radius for self-fusion. In Figure 2.4(a), the target b-scan frame $\mathbf{s}_t$ is highlighted with red border. The Network I is trained by the mean square error

(MSE):

$$\mathscr{L}_{MSE}(\mathbf{s}_t, \tilde{\mathbf{s}}_t) = \frac{1}{HW}\sum_{i}^{H}\sum_{j}^{W}(\tilde{\mathbf{s}}_t(i,j) - \mathbf{s}_t(i,j))^2 \tag{2.10}$$

where i, j are the index of height H and width W of the image, $\tilde{\mathbf{s}}_t(i,j)$ and $\mathbf{s}_t(i,j)$ are the intensity values at pixel (i,j). $\tilde{\mathbf{s}}_t$ is the resultant pseudo-modality created.

2.6.1.2 Network II: Pseudo-multimodal fusion

The noisy b-scan $\mathbf{x}_t$ has fine details including small vessels and texture, while the speckle noise is too strong to clearly reveal layer structures. The pseudo-modality $\tilde{\mathbf{s}}_t$ has well-suppressed speckle noise and clean layers, but many of the subtle features are lost. Therefore, merging the essential features from these mutually complementary modalities is my goal. To produce an output that inherit features from two sources, Network II takes feedback from the ground truth of both modalities in seeking for a balance between them. I use L1 loss for $\mathbf{y}_t$ to punish loss of finer features and MSE for $\mathbf{s}_t$ to encourage some blur effect in layers. The weight of these loss functions are determined by hyper-parameters. The overall loss function is:

$$\mathscr{L} = \frac{\alpha}{HW}\sum_{i}^{H}\sum_{j}^{W}|\tilde{\mathbf{y}}_t(i,j) - \mathbf{y}_t(i,j)| + \frac{\beta}{HW}\sum_{i}^{H}\sum_{j}^{W}(\tilde{\mathbf{y}}_t(i,j) - \mathbf{s}_t(i,j))^2 \tag{2.11}$$

where parameters α and β are the weights of the two loss functions, and they can be tuned to reach a trade off between layers from the pseudo-modality and the small vessels from the HN b-scan.

Because of the poor quality of single frame b-scan, more supplementary information and constraints are likely to be beneficial for feature preservation. For instance, observing the layered structure of the retina, Ma et al. (2018) introduce an edge loss function to preserve the prevailing horizontal edges. Devalla et al. (2019) investigate a variation to U-Net architecture so that the edge feature is enhanced. To emphasize the edge information, I implement the 3 × 3 Sobel kernels to compute the image gradient $\mathbf{g}_t$ from the pseudo-modality $\tilde{\mathbf{s}}_t$. Figure 2.4(b) shows that Network II thus takes a three-channel input.

2.6.2 Experiments

In this study, my goal is to show that the denoising result is improved by the processing pipeline that introduces the pseudo-modality. Thus, I do not focus on varying the network structure for better performance. Instead, I will use the Network II with single channel input $\mathbf{x}_t$ as the baseline. For this baseline, the loss function will only have feedback from $\mathbf{y}_t$. I hypothesize that the relative results between single modality and pseudo-multimodal denoising will have a similar pattern for other architectures for Network II, but exploring this is beyond the scope of the current study. Since the network architecture is not the focus of my study, I use the same multi-scale U-Net (MSUN) architecture proposed by Devalla et al. (2019), for both Network I

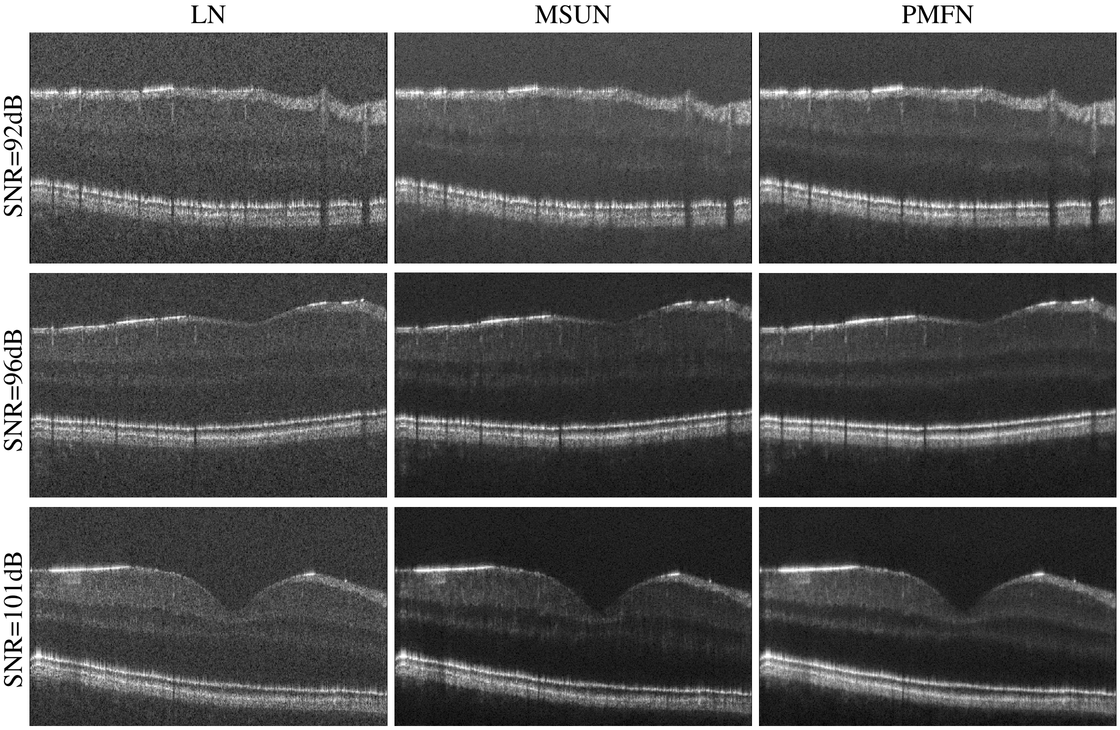


Figure 2.5: Fovea denoising results for different input SNR (Excess background trimmed)

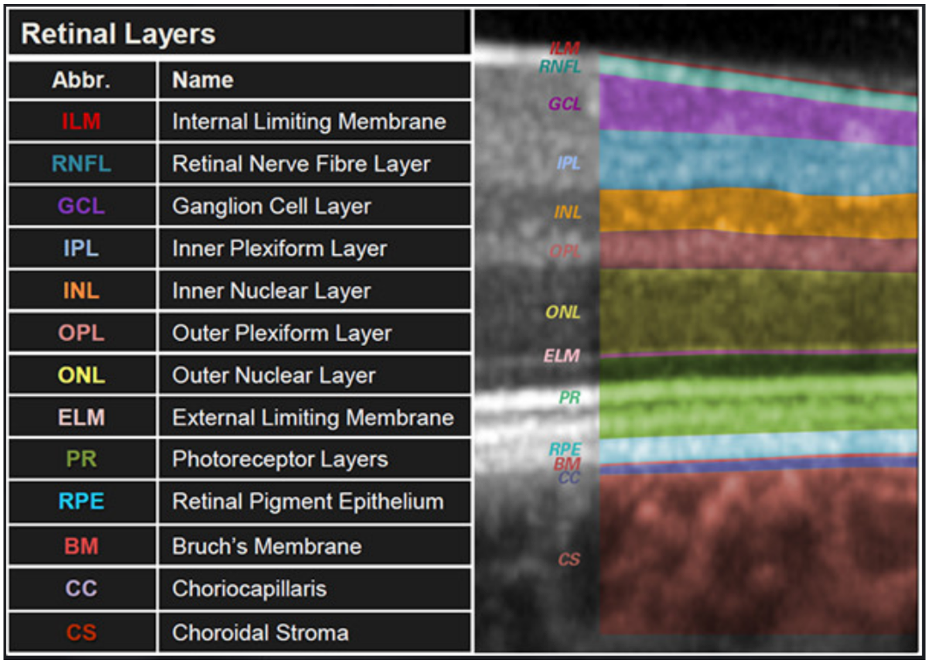
and II.

The b-scan neighborhood radius for self-fusion was set at $R = 7$. All the models are trained on NVIDIA RTX 2080TI 11GB GPU for 15 epochs with batch size of 1. Parameters in network are optimized by Adam optimizer with starting learning rate 10^{-4} and a decay factor of 0.3 for every epoch. In Network II, I use $\alpha = 1$ and $\beta = 1.2$.

2.6.3 Results

2.6.3.1 Qualitative results

Figure 2.6 denotes the human retina layers in an OCT-A b-can. In the qualitative evaluation, I focus on the boundaries between the ganglion cell layer (GCL), the inner plexiform layer (IPL), the inner nuclear layer (INL) and the outer plexiform layer (OPL).



Retinal Layers

Abbr.	Name
ILM	Internal Limiting Membrane
RNFL	Retinal Nerve Fibre Layer
GCL	Ganglion Cell Layer
IPL	Inner Plexiform Layer
INL	Inner Nuclear Layer
OPL	Outer Plexiform Layer
ONL	Outer Nuclear Layer
ELM	External Limiting Membrane
PR	Photoreceptor Layers
RPE	Retinal Pigment Epithelium
BM	Bruch's Membrane
CC	Choriocapillaris
CS	Choroidal Stroma

Figure 2.6: Retinal layers in an OCT b-scan. Image adapted from https://www.heidelbergengineering.com/int/news/know-your-retinal-layers-33401465/

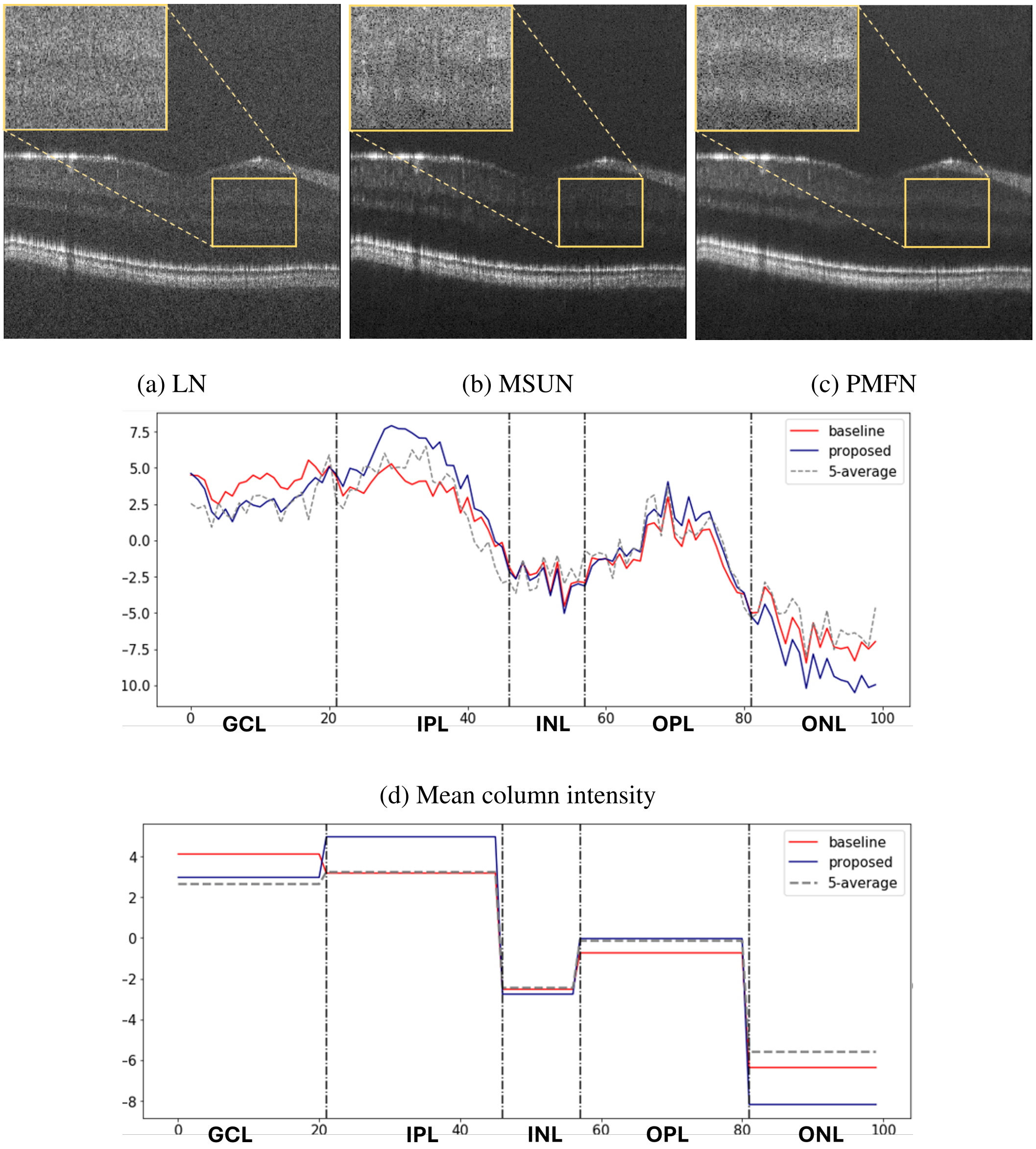
Figure 2.5 displays the denoising performance of the proposed algorithm for different input SNR levels. Compared to the baseline model, I observe that PMFN has better separation between GCL and IPL, which enables the vessels in GCL to better stand out from noise. Moreover, the improvement of smoothness and homogeneity in OPL makes it look more solid and its border more continuous. In addition, the retinal pigment epithelium (RPE) appears to be more crisp.

In Figure 2.7, to better assess the layer separation, I focus on a b-scan with high speckle noise (SNR=92) that severely obscures the boundary between layers. In the top row (a-c), I zoom into a region of interest (ROI) that contains 5 tissue layers (from top to bottom): GCL, IPL,INL, OPL and outer nuclear layer (ONL). As the baseline model learns only from the high noise b-scan, layer boundaries are not clear: GCL and IPL are indistinguishable, and although the INL and OPL are preserved, they are not as homogeneous as in the PMFN result. PMFN remedies these problems.

Another way of assessing the separability of layers or, in other words, the contrast between adjacent layers, is plotting the column intensity (Figure 2.7(d)). Since the layers within the ROI are approximately flat, I take the mean vector along the row. In order to rule out any potential difference of intensity level, I normalize the mean vector with the average intensity of ROI. The mean vector is thus given by

$$\bar{\mathbf{v}} = \frac{1}{W_{ROI}} \sum_{i}^{W_{ROI}} \mathbf{v}_i - \mu_{ROI} \tag{2.12}$$

where W_{ROI} is the width of the ROI, $\mathbf{v}_i$ is a column vector in the window and μ_{ROI} is a vector that has the mean of the ROI as all its elements. I plot the $\bar{\mathbf{v}}$ for Figure 2.7-a, Figure 2.7-b and Figure 2.7-c in Figure 2.7-d. The border between layers are approximated with vertical dash lines for this visualization. In Figure 2.7-d, the proposed method tends to have lower intensity in dark bands and higher intensity in bright ones. This indicates that it has better contrast between adjacent layers. Figure 2.7-e summarizes the mean intensity within each layer. Because of high intensity speckle noise, the baseline result completely misses the GCL-IPL distinction, whereas my method provides good separation.



(a) LN

(b) MSUN

(c) PMFN

(d) Mean column intensity

(e) Mean layer intensity

Figure 2.7: Layer separation analysis

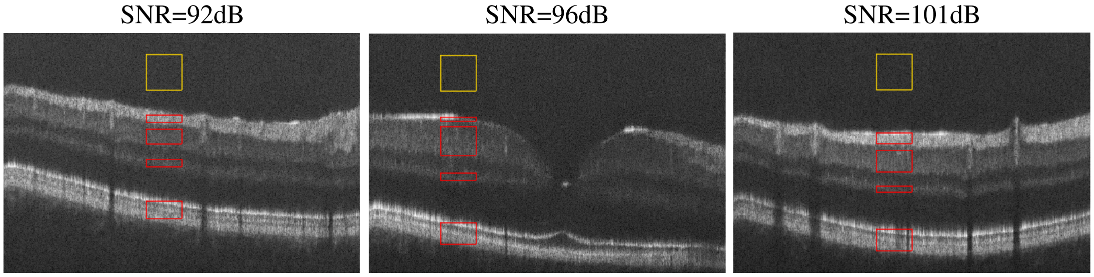


Figure 2.8: Background (yellow) and foreground (red) ROIs

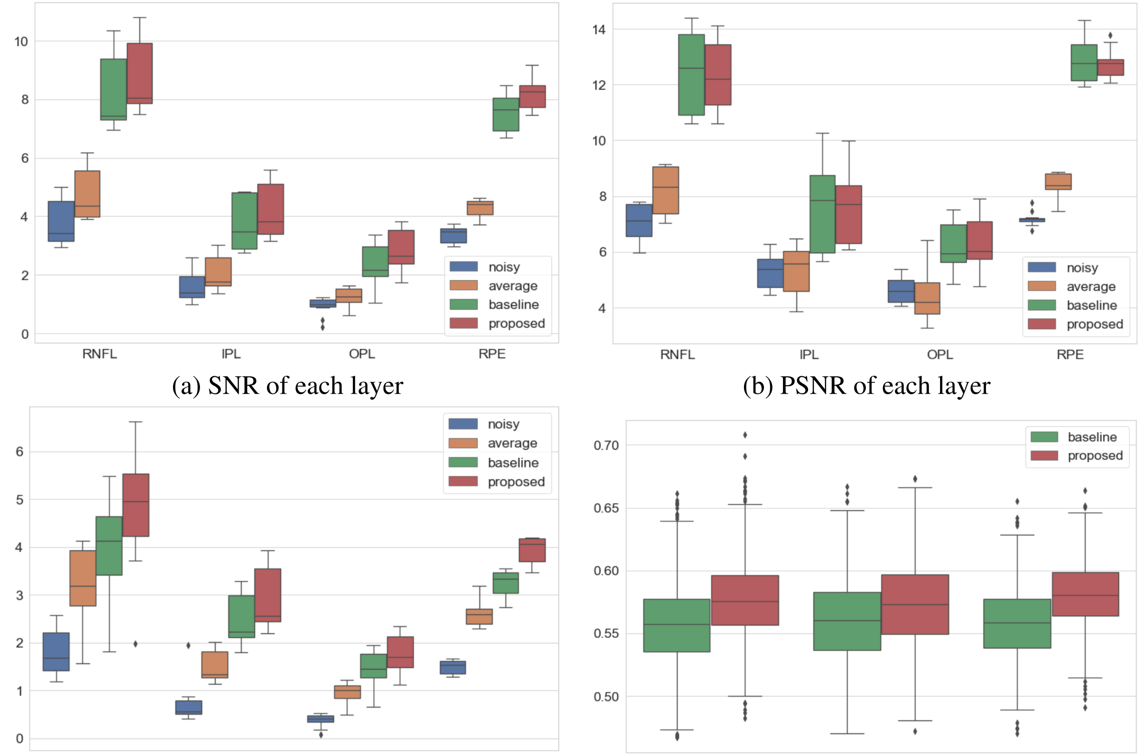
2.6.3.2 Quantitative results

I report the signal-to-noise ratio (SNR), peak signal-to-noise ratio (PSNR), contrast-to-noise ratio (CNR) and structural similarity (SSIM) of my results. I used the 5-frame-average LN image as the ground truth although it is far from being noiseless.

As it is illustrated in Figure 2.6, every retinal layer has a different intensity level, so I report each metric separately for RNFL, IPL, OPL and RPE. I manually picked foreground and background ROIs from each layer, as shown in Figure 2.8, for 10 b-scans. To avoid local bias, these chosen slices are far apart to be representative of the whole volume. When computing metrics for a given layer, the background ROI (yellow box) is cropped as needed to match the area of the foreground ROI (red box) for that layer.

Figure 2.9 (a) to (c) display the evaluation result for SNR, PSNR and CNR respectively. For all layers, the proposed model gives the best SNR and CNR results, while the PSNR stays similar with the baseline multi-scale UNet model. The proposed method also shows the improvement in terms of SSIM compared with the baseline.

This study shows that the self-fusion pseudo-modality can provide major contributions to OCT denoising by emphasizing tissue layers in the retina. The fusion network allows the vessels, texture and other fine details to be preserved while enhancing the layers. Although the inherent high dimensionality of the deep network has sufficient complexity, more constraints in the form of additional information channels are able to help the model converge to a desired domain. This work is published on Ophthalmic Medical Image Analysis: 7th International Workshop Hu et al. (2020) and the source code and model checkpoint are publicly available at https://github.com/DeweiHu/Real-time-PMFN. My collaborator Rico-Jimenez et al. (2022) implement my model on the real-time OCT imaging system.



(a) SNR of each layer

(b) PSNR of each layer

(c) CNR of each layer

(d) SSIM for input of different noise level

Figure 2.9: Quantitative evaluation of denoising results

2.7 Unsupervised-learning Method: DDPM

As stated before, the traditional way to reduce speckle noise is to average multiple noisy b-scan acquisitions at the same location. This approach requires a large number of repetitions, and the prolonged acquisition time can be problematic for patient comfort; additionally, registration artifacts caused by eye movement can be an issue. Therefore, a denoising algorithm that does not require repeated acquisitions is desirable. In recent research, Ho et al. (2020) proposed a diffusion denoising probabilistic model (DDPM) that is used for image synthesis with a performance superior to generative adversarial networks (GANs) (Dhariwal and Nichol, 2021). The general idea of the diffusion model is straightforward: Given a natural image, a Markov chain can be formed by adding a small amount of Gaussian noise at each step. After a sufficient number of steps, a complex data distribution will finally be transformed to a Gaussian. Conversely, given a Gaussian noise image, a meaningful image can be synthesized in a reverse process. In this work, I propose to leverage the diffusion probabilistic model to denoise retinal OCT b-scans. Since the model is learning the speckle pattern instead of the retina appearance, the reference image used for training is not required to be the true noise-free image. In my experiments I apply the self-fusion method to obtain the clean reference image and I train the parameterized Markov chain by variational inference. As the number of reverse steps is adjustable, the algorithm is able to produce different levels of denoising results. This is an advantage since different tasks may require different levels of fine detail retention in the images. Compared with the PMFN described in Section 2.6, this approach does not require the low-noise image in training. In the result evaluation I show that the DDPM has better performance regarding the SNR, CNR and ENL.

2.7.1 Method

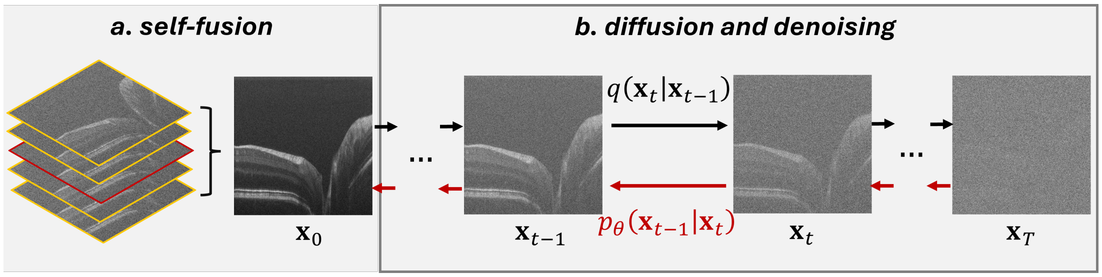

a. self-fusion
b. diffusion and denoising
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_0$
$\mathbf{x}_{t-1}$
$\mathbf{x}_t$
$\mathbf{x}_T$
...
...

Figure 2.10: General workflow for the diffusion probabilistic model.

I leverage my self-fusion as a pre-processing step in the training stage (Figure 2.10a). As previously discussed, this approach is easy to implement and robust for retinal layer enhancement, but finer features like vessels and texture can be over-smoothed. Nevertheless, the diffusion probabilistic model aims to learn the speckle pattern instead of the signal, the self-fusion output $\mathbf{x}_0$ can still be used as the clean image for my training purposes. Figure 2.10b shows a Markov chain in forward (diffuse) and reverse (denoise) directions:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{2.13}$$

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{2.14}$$

where $q(\mathbf{x}_0)$ represents the data distribution while $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_t; \mathbf{0}, \mathbf{I})$. θ is the parameters of the model. The diffusion and sampling describe a transition between these two distributions with T discretized steps. My goal is to train a deep model p_θ to restore a noisy image $\mathbf{x}$ with an adjustable parameter t. Intuitively, an image with stronger speckle will require a larger t value that indicates more denoising steps.

The image sequence $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T$ is created by gradually adding small Gaussian noise with a variance schedule $\{\beta_1, \ldots, \beta_T\}$, where $\beta_t \in (0,1)$, $\forall t \in (1,T)$.

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad \text{where} \quad \alpha_t = 1 - \beta_t \tag{2.15}$$

Eq. 2.15 approximates the posterior distribution in the forward process assuming that there is a small mean shift after one step of diffusion. Denote $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, then $\mathbf{x}_t$ is obtained by adding t different Gaussian random variables to $\mathbf{x}_0$.

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2.16}$$

As the sum of t Gaussians is a Gaussian with variance $\sum_{t=1}^{T} \beta_t \approx 1 - \bar{\alpha}_t$. The high order terms with regard to β_t in $1-\bar{\alpha}_t$ are negligible because β_t is a small value in range $(0,1)$. In practice, Eq. 2.16 enables sampling of $\mathbf{x}_t$ with reparameterization:

$$\mathbf{x}_t(\mathbf{x}_0, \varepsilon) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \quad \varepsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2.17}$$

Similar to Eq. 2.15, the reverse step is also modeled as a Gaussian since the noise added in each step is small:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \tag{2.18}$$

In this work, I set the variance to be a fixed parameter $\Sigma_\theta(\mathbf{x}_t, t) = \beta_t$ and only learn to predict the mean. To perfectly recover the image in the reverse process, the ideal solution is to minimize the distance between $q(\mathbf{x}_{t-1}|x_t)$ and $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$. However, according to Ho et al. (2020), the direct KL-divergence between these two distributions, $D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$, is not tractable. Alternatively, they introduce a tractable constraint $D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$. By leveraging the property of Markov chain, the following equation should hold:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} = q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0) \tag{2.19}$$

From Eq. 2.15 and Eq. 2.16, I know that both terms in the product are Gaussian; then $q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)$ should also have the form $\mathscr{N}(\mathbf{x}_{t-1};\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0),\tilde{\beta}_t\mathbf{I})$:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) = \mathscr{N}\left(\mathbf{x}_{t-1};\underbrace{\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0+\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t(\mathbf{x}_0,\varepsilon)}_{\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0)},\underbrace{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}_{\tilde{\beta}_t}\mathbf{I}\right), \quad \varepsilon \in \mathscr{N}(\mathbf{0},\mathbf{I}) \tag{2.20}$$

I use the negative log likelihood $-\log p_\theta(\mathbf{x}_0)$ as the objective function. This can be optimized by minimizing its variational upper bound $\mathscr{L}$ given by the Jensen's inequality.

$$\mathscr{L} = \underbrace{D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))}_{\mathscr{L}_T} + \sum_{t=2}^{T}\underbrace{D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{\mathscr{L}_{t-1}} + \underbrace{\mathscr{H}(p_\theta(\mathbf{x}_0|\mathbf{x}_1))}_{\mathscr{L}_0} \tag{2.21}$$

where $\mathscr{H}$ denotes entropy. Because the variance schedule is fixed in this implementation, $\mathscr{L}_T$ turns out to be a constant, so I only need to consider $\mathscr{L}_{t-1}$ and $\mathscr{L}_0$ as loss function. Given Eq. 2.20 and Eq. 2.18, $\mathscr{L}_{t-1}$ is the KL divergence of two Gaussian distributions and can be reduced to Eq. 2.22.

$$\mathscr{L}_{t-1} = \frac{1}{2\beta_t}\left\|\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0) - \mu_\theta(\mathbf{x}_t(\mathbf{x}_0,\varepsilon),t)\right\|^2 \tag{2.22}$$

Obviously, to minimize $\mathscr{L}_{t-1}$ I can set the mean prediction equal to $\tilde{\mu}_t(x_t,x_0)$ which is derived from Eq. 2.20 and Eq. 2.17:

$$\mu_\theta(x_t,t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(\mathbf{x}_t,t)\right) \tag{2.23}$$

Combining Eq. 2.22 and Eq. 2.23 it is easy to see that the model is learning to predict the sample ε drawn from a normal distribution with mean square error (MSE). The expression of $\mu_\theta(\mathbf{x}_t,t)$ demonstrates that the image restoration is actually done by adding a Gaussian noise.

2.7.2 Experiments

For self-fusion, I take adjacent b-scans within radius of $R = 3$ as candidates to compute the weighted sum. Images are padded to 512×512 pixels, and the intensity is normalized to $[-1,1]$. The variance schedule for the β_t is set to linearly increase from 10^{-4} to 6×10^{-3} in $T = 100$ steps. The architecture of my model is a residual U-Net. It is trained on an NVIDIA RTX 2080TI 11GB GPU for 500 epochs, with a batch size of 2 using Adam optimizer. The starting learning rate is 10^{-4} and decays by half every 5 epochs.

2.7.3 Results

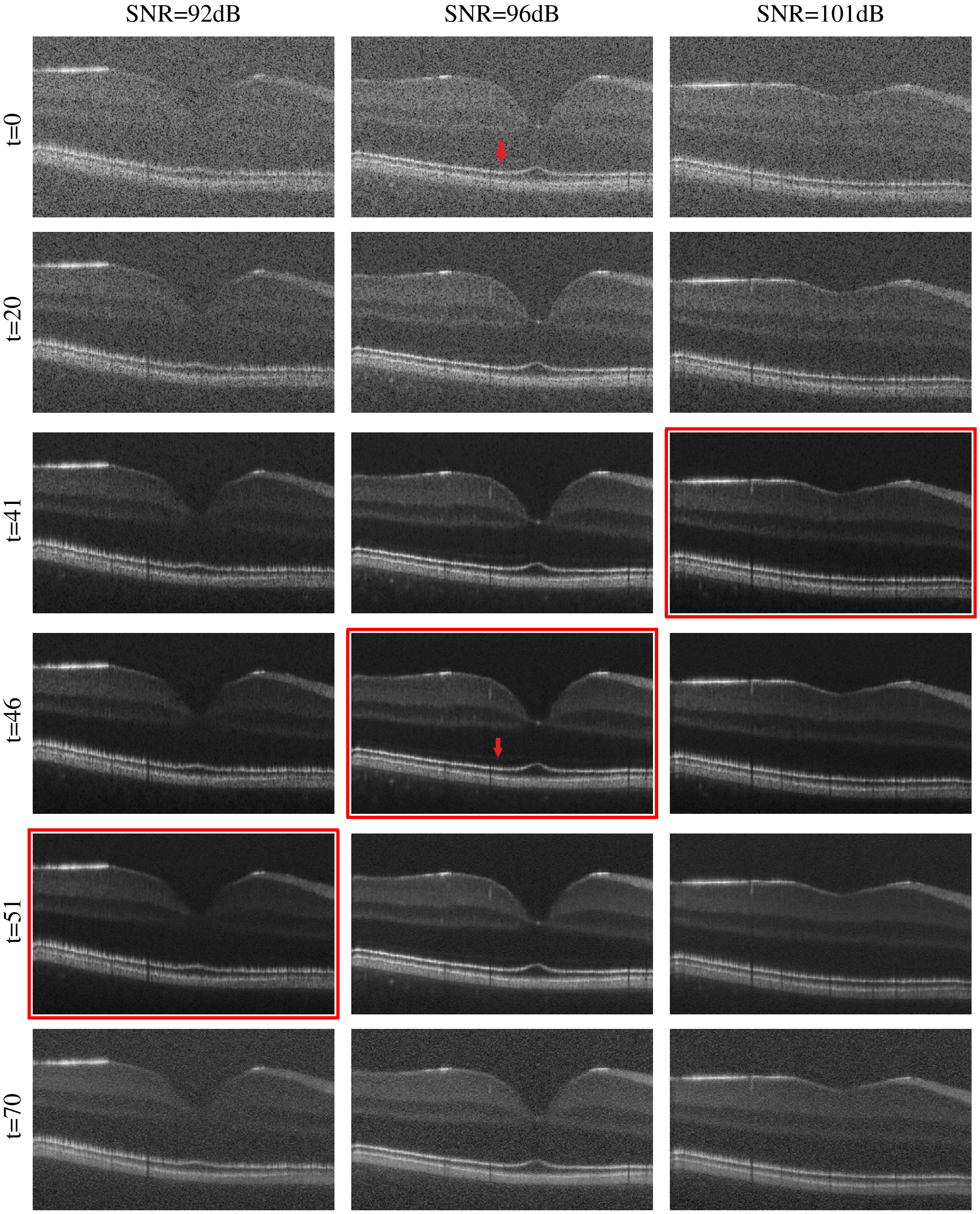


Figure 2.11: Fovea denoising results for different input SNR levels and for different t values

In Figure 2.11 I show the denoising results of my model for a range of t values. $t = 0$ is the original input image for each SNR level. Increasing t values indicate more denoising steps. The best result (determined visually) for each noise level, as highlighted by the red box, coincide with the intuition that the noisier images benefit from a larger t. For the third column, where the input noise level is relatively low, obviously as t increases from 41 to 51, retinal layers gradually become over-smoothed and fine texture features fade away. As discussed before, Eq. 2.23 explains that the denoising process is done by adding Gaussian noise to compensate for the speckle pattern. When the t is too large (e.g., $t = 70$ in Figure 2.11), the added noise becomes excessive and produces poor results. I further note that my proposed method performs well for vessel and layer preservation. For example, in the second column, my result reveals the very thin external limiting membrane (ELM) (marked by red arrows) which is hardly visible in the noisy input. In practice, the optimal t is determined by the ENL of the background. In Figure 2.12, I show an example for the volume with $SNR = 96dB$. The optimal value for denoising step is $t = 46$.

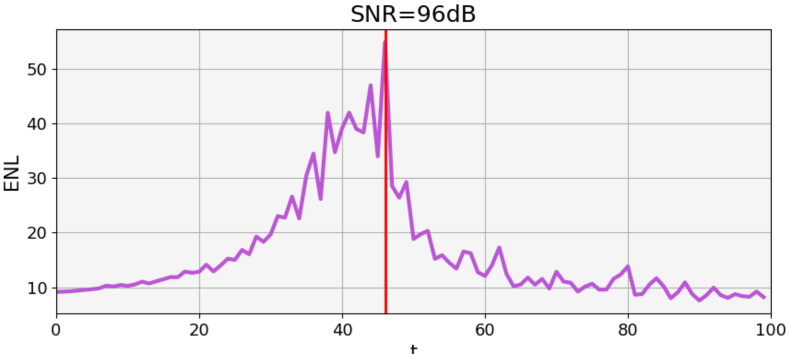


Figure 2.12: Background ENL for different t values.

In the previous section, I presented a pseudo-modality fusion network (PMFN) that improves the feature preservation compared to a former deep learning method developed by Devalla et al. (2019). Here, I compare with the PMFN to show that the denoising probabilistic model has superior denoising performance. In Figure 2.13, I observe that the retinal layers are more homogeneous in my proposed approach than in PMFN for all input SNR levels. Downstream analysis tasks such as layer segmentation would likely benefit from this improvement. I also note that small features like vessels are not sacrificed, even though other regions of the layers become denoised.

To quantitatively confirm these observations, I use the average of 5 repeated frames (5-mean) as the reference ground truth image and I report several metrics in Table 2.1. Comparing with PMFN, the proposed method improve the denoising performance in terms of signal-to-noise ratio (SNR), peak signal-to-noise ratio (PSNR), contrast-to-noise ratio (CNR) and equivalent number of looks (ENL). Improvements over the baseline are highlighted in boldface. The results are significantly different in a paired, two-tailed t-test with a significance threshold of 0.05.

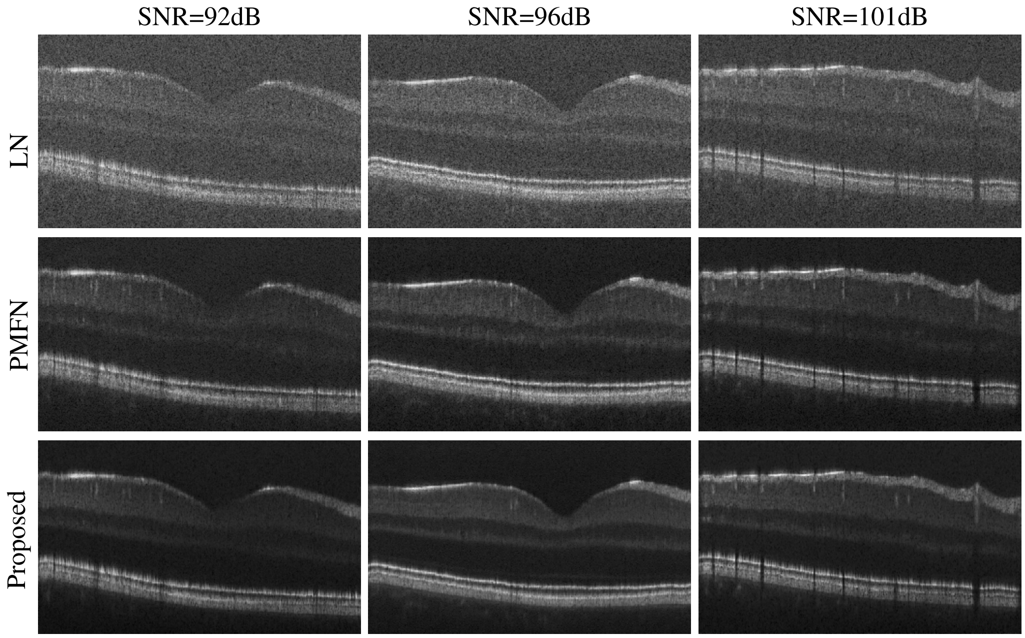


Figure 2.13: Result comparison with baseline model.

	SNR	PSNR	CNR	ENL
Hybridly supervised	29.18 ± 2.03	81.51 ± 0.69	1.89 ± 0.57	10.91 ± 2.80
Unsupervised	**40.94 ± 1.78**	74.67 ± 0.58	**2.12 ± 0.71**	**54.66 ± 15.84**

Table 2.1: Quantitative evaluation

The model is unsupervised and requires a small amount of data to train. Moreover, the parameter t provides control over the output smoothness level. My results show that my model can efficiently suppress the speckle; furthermore, the features including layers and vessels are not only preserved but present higher visibility. A limitation of this approach is its Gaussian assumption on the speckle pattern. The generalization to other noise distribution types can be a potential direction of future work for better speckle modeling. This work is published on *Medical Imaging 2022: Image Processing* Hu et al. (2022b) and the source code and model checkpoint are publicly available at https://github.com/DeweiHu/OCT_DDPM.

2.8 Chapter summary

Base on the non-learning-based self-fusion algorithm, I propose a supervised learning method (PMFN) and an unsupervised approach (DDPM) for OCT b-scan noise reduction. In general, both methods significantly improve the image quality. Being the main obstacle for the implementation of deep learning in medical image processing, the need for ground truth is gradually attenuated with the development of unsupervised approaches. Although the algorithm development is progressing fast, numerically evaluating the denoising results in the absence of noise-free reference is still a non-trivial problem. None of the existing popular evaluation metrics are not accurate enough, especially for small features in the retina. More importantly, their definition vary in different works, which makes it hard to directly compare the performance of the methods. Hence, the quality control and benchmarking can be one of the essential areas of future research.

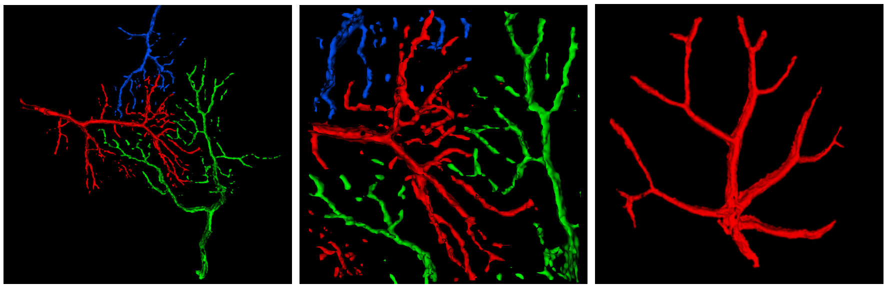
CHAPTER 3

Enhanced Angiogram Synthesis

3.1 Introduction

As indicated in Chapter 1, the retinal vessels are usually visualized in OCT angiography (OCT-A), in which the vasculature with dynamic blood flow is decoupled from the surrounding stationary tissue. The OCT-A is popular for studying choroidal neovascularization (Burke et al., 2017), diabetic retinopathy (Ishibazawa et al., 2015) and other retinal pathologies. Recent findings that the vascular plexus density is an important indicator for the severity of certain eye diseases (Holló, 2018) have further highlighted the need for density computation via vessel segmentation in OCT-A. Therefore, an algorithm that can provide accurate vessel segmentation from the OCT-A would have profound impact on future clinical practice. In recent years, deep learning models have achieved remarkable success in this task. Nevertheless, most of the related research is conducted on 2D depth projection of OCT-A volumes (Giarratano et al., 2019; Ma et al., 2020; Liu et al., 2022). To be clear, this does not refer to a scenario where each 2D slice of a 3D volume is segmented independently, which would allow the resulting 2D segmentation maps to be stacked together to create a 3D segmentation in post-processing, as is common practice in MRI segmentation. Such an approach remains elusive for OCT-A due to the very poor SNR and lack of training data. Rather, the solution consists of collapsing the whole 3D volume into a single 2D image (e.g., the depth projection shown in Figure 1.4), such a projection image which has better SNR and can be manually segmented. This approach only produces a single 2D segmentation out of a whole 3D volume, and evidently sacrifices the depth information that would have been present in a 3D analyses.

To extend the application of deep learning vessel segmentation to volumetric data, the primary obstacle to break through is the lack the ground truth. To the best of my knowledge, there is no manually annotated public dataset for 3D OCT-A vasculature. As a result, using an unsupervised or self-supervised learning algorithm is advisable for tackling this issue. Typically, the process involves training a conditional generative model to improve the visibility of the vessels in angiographic images, followed by the application of a simple thresholding technique to binarize the output. As an example, Zhang et al. (2019) apply the model-based optimal oriented flux (Law and Chung, 2008) to enhance the 3D vessels and threshold the result for segmentation on OCT-A. In my research, I propose a novel method to synthesize the enhanced angiogram for OCT-A and fundus photography (FP). This not only facilitates the development of self-supervised 3D vessel segmentation but also produces various 2D pseudo-modalities featuring prominently enhanced vessels. These



(a) human retina label (b) cropped ROI (c) zebrafish retina label

Figure 3.1: Three manually labelled vessel plexuses (shown in red, green and blue). Only the branches contained in the cropped ROI (panel b) are fully segmented; any branches outside the ROI and all other trees are omitted for brevity. Panel c is the 3D rendering of the zebrafish retinal vessel. I label the whole volume manually since the structure is very simple.

pseudo-modalities are implemented as augmented data in my following research about domain generalization presented in Chapter 4.

In the following sections, I first introduce the preliminary synthetic model utilized for angiogram enhancement in Section 3.3. Then I showcase its implementation on both 3D OCT-A and 2D fundus photography data in Section 3.4 and Section 3.5, respectively.

3.2 Datasets

3D OCT-A data. The original OCT volumes were acquired in 4 seconds with $2560 \times 500 \times 400 \times 4$ pix. (spectral × lines × frames × repeated frames) with a 2.5 ms interscan delay (Malone et al., 2019; El-Haddad et al., 2018). These images are acquired with a handheld OCT system, which leads to an increased amount of motion artifacts. Lateral and axial bulk-motion were removed by a discrete Fourier transform based registration (Guizar-Sicairos et al., 2008) on sequential OCT images. OCT-A was performed on the motion-corrected OCT volumes using singular value decomposition. In this work, I concentrate on segmenting vasculature within the ganglion cell layer (GCL) and inner plexiform layer (IPL), which contain most of the vessels near the fovea. I manually crop the volume to only retain the depth slices within GCL and IPL. Three fovea volumes are used for training and one kept for testing. As only a limited number of slices exist between GCL and IPL, I aggressively augment the dataset by arbitrarily cropping and randomly flipping 10 windows of size $[320, 320]$ for each *en-face* image. To assess the segmentation performance on vessels differing in size, I manually labeled 3 interacting plexus near the fovea, which are rendered in Figure 3.1a. A smaller ROI ($120 \times 120 \times 17$) cropped in the center (Figure 3.1b) is used for numerical evaluation.

Other than human retinal OCT-A, I use OCT-A of zebrafish eyes which have a simple vessel structure ideal for easy manual labeling (Figure 3.1c). This also allows me to test the generalizability of my method

dataset	modality	resolution	number
DRIVE (Staal et al., 2004)	fundus	565×584	20
STARE (Hoover et al., 2000)	fundus	700×605	20
ARIA (Farnell et al., 2008)	fundus	768×576	$61/59/23$

Table 3.1: 2D FP datasets for pseudo-modality generators training. Note that the ARIA has class labels for disease status, thus I list the number of samples for classes healthy/diabetic/age-related macular degeneration (AMD).

to images from different species. Furthermore, the zebrafish dataset contains stronger speckle noise than the human data, which can help to test the robustness of the method to high noise. 3 volumes ($480 \times 480 \times 25 \times 5$ each) are labeled for testing and 5 volumes are used for training. All the OCT-A datasets are provided by DIIGI Lab at Vanderbilt University and the manual labelling is done on ITKSnap (Yushkevich et al., 2006) by myself.

2D fundus photography data. I use three annotated public FP datasets to train the pseudo-modality generators. The details about these datasets are listed in Table 3.1.

3.3 Preliminary: Vessel Enhancement Model

Liu et al. (2020) proposed an unsupervised 2D vessel segmentation method with a variational intensity cross channel encoder (VICCE). Their approach implements two registered OCT-A depth projection created by two different scanning devices (Spectralis and Cirrus) taken on the same subject. A variational autoencoder works as a pix2pix (Isola et al., 2017) translator to map one image to the other, and the latent space is set to have the same dimension with the input. Here, I denote the source image as $\mathbf{x}^s \in \mathbb{R}^{C \times H \times W}$ and the target image as $\mathbf{x}^t \in \mathbb{R}^{C \times H \times W}$, where C is the number of channels, and H and W are the height and width of the image. Then the latent representation is set to be a grayscale image $\mathbf{z} \in \mathbb{R}^{H \times W}$.

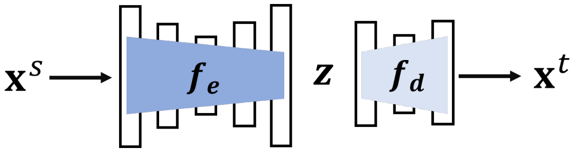


Figure 3.2: Synthesis model structure. f_e and f_d are the encoder and the decoder. Both models have the U-Net architecture and f_e has more U-Net layers than f_d.

As shown in Figure 3.2, the synthesis model has a encoder-decoder structure. Both f_e and f_d have the U-Net (Ronneberger et al., 2015) architecture. The desired output of the model is the latent image $\mathbf{z} = f_e(\mathbf{x}^s)$. Therefore, more learnable parameters are assigned to f_e such that it has stronger ability to learn vessel related features from the source image $\mathbf{x}^s$. The decoder is used to provide supervision during training with a reconstruction loss:

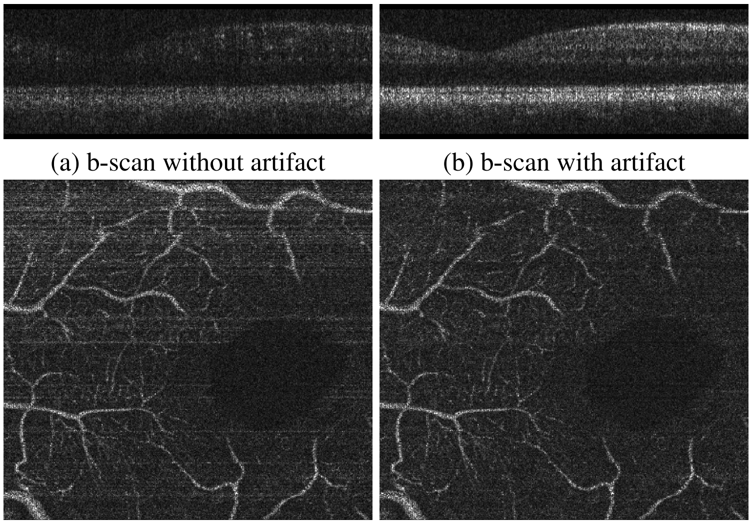
$$\mathscr{L}_{reconst} = \frac{\alpha}{HW}\sum_{i}^{H}\sum_{j}^{W}|\tilde{\mathbf{x}}^{t}(i,j)-\mathbf{x}^{t}(i,j)| + \frac{\beta}{HW}\sum_{i}^{H}\sum_{j}^{W}\left(\tilde{\mathbf{x}}^{t}(i,j)-\mathbf{x}^{t}(i,j)\right)^{2} \tag{3.1}$$

where $\tilde{\mathbf{x}}^{t} = f_d(f_e(\mathbf{x}^s))$, i and j are the pixel position indicators and alpha and beta are the hyper-parameters to adjust the weight of the L1 norm and the MSE. According to previous works in unsupervised segmentation (Liu et al., 2020) and representation disentanglement (Dewey et al., 2020; Ouyang et al., 2021), this model is able to effectively extract interpretable visual representations that have direct spatial correspondence with the original image. In the context of causal representation learning, if I regard the images $\mathbf{x}^s$ and $\mathbf{z}$ are two sets of features, $\mathbf{z}$ must include all the causal features to fully reconstruct $\mathbf{x}^t$. Given that $\mathbf{x}^s$ and $\mathbf{x}^t$ are two different OCT-A taken from the same subject, the vessel structure can, thus, be regarded as the causal feature as they are supposed to be identical. Consequently, to successfully reconstruct the target image, $\mathbf{z}$ should contain all the vessels such that $P(\mathbf{x}^t|\mathbf{z}) \approx P(\mathbf{x}^t|\mathbf{x}^s)$. The conditional probability $P(\mathbf{x}^t|\mathbf{z})$ is modeled by the decoder f_d. Experimentally, I observe that the encoder f_e can filter out irrelevant features (e.g., speckle noise) while greatly enhancing the shared features extracted from the source image. Therefore, a simple thresholding on the synthesized $\mathbf{z}$ can provide a decent binary vessel mask without any ground truth for training.

Unfortunately, the major obstacle that prevents the wide adoption of this method (in its original form, as proposed by Liu et al. (2020)) is: imaging the same retina with different devices is rarely possible even in research settings and unrealistic in clinical practice. However, by adjusting the setting of source image $\mathbf{x}^s$ and target image $\mathbf{x}^t$, this method can still be implemented for enhanced angiogram generation. In the following sections, I introduce the application of this synthesis model in volumetric OCT-A and 2D FP data.

3.4 3D Vessel Segmentation

In this section, I extend the 2D vessel segmentation to 3D by segmenting slice by slice in the depth axis (dubbed as *en-face* images) of the OCT-A volume. Usually, blood flows faster at the center of a large vessel, which appears brighter in OCT-A. In contrast, small capillaries often have low intensity and are prone to quickly vanish with the change of depth. Consequently, they are hardly distinguishable from the ubiquitous speckle noise. I exploit the similarity of vasculature between consecutive *en-face* OCT-A slices to improve the image quality. This is achieved by the aforementioned self-fusion (Oguz et al., 2020) introduced in Chapter 2. Then the self-fusion image will be utilized as the target image $\mathbf{x}^t$ in the synthesis model as it has inherent pixel-wise correlation with the input *en-face* image. The proposed method is named as local intensity fusion encoder (LIFE).



(a) b-scan without artifact

(b) b-scan with artifact

(c) original *en-face* image

(d) cleaned *en-face* image

Figure 3.3: Original *en-face* image and artifact removal result. The horizontal bright strips in the original *en-face* image (c) correspond to the ill-decoupled OCT-A b-scan in (b). The artifact removal effectively suppresses these lines (d).

3.4.1 Method

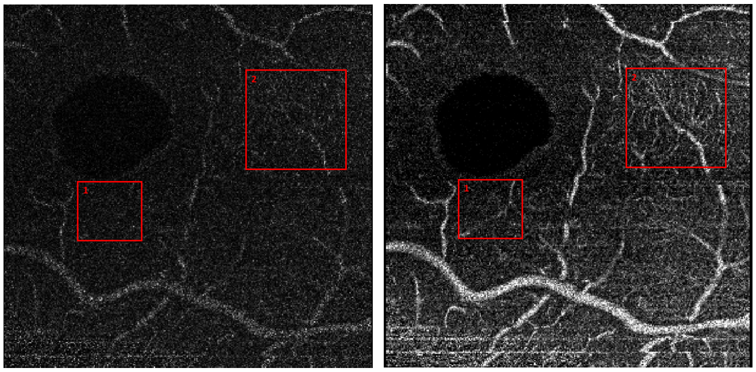
3.4.1.1 Pre-processing

Decorrelation allows OCT-A to emphasize vessels while other tissue types get suppressed (Figure 3.3a). However, this requires the repeated OCT b-scans to be precisely aligned. Any registration errors between b-scans cause motion artifacts, such that stationary tissue is not properly suppressed (Figure 3.3b). These appear as horizontal artifacts in *en-face* images (Figure 3.3c). The OCT-A b-scans with motion artifacts can be easily detected by simply thresholding on the mean image intensity. They have much higher mean value than those without artifact. I then remove these artifacts by matching the histogram of the artifact B-scan to its closest well-decorrelated neighbor (Figure 3.3d). This solution is simple while effective since I work on *en-face* images for this specific task.

3.4.1.2 Enhanced angiogram synthesis

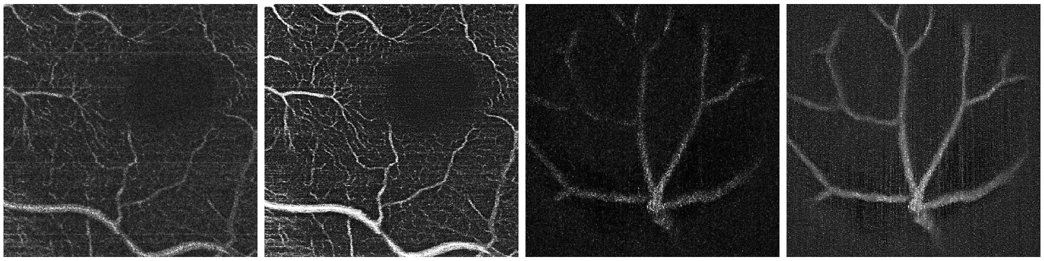
Instead of conducting self-fusion on b-scans, for each 2D *en-face* slice $\mathbf{x}_i$ (the depth indexed by the subscript i), I compute self-fusion from its adjacent slices within an R-neighborhood $\{\mathbf{x}_{-R}, \dots, \mathbf{x}_{+R}\}$. In order to make vessels stand out better, I enhance the contrast of the self-fusion result with the built-in function (https://pillow.readthedocs.io/en/stable/reference/ImageEnhance.html) to get $\mathbf{s}_i$. Figure 3.4 compares $\mathbf{x}_i$ and $\mathbf{s}_i$.

However, self-fusion of *en-face* images sacrifices the accuracy of vessel diameter in the depth direction. Specifically, some vessels existing exclusively in neighboring images are inadvertently projected on the target slice. For example, the small red box in Figure 3.4 highlights a phantom vessel caused by incorrect fusion.



(a) original *en-face* slice $\mathbf{x}_i$ (b) contrast-enhanced self-fusion $\mathbf{s}_i$

Figure 3.4: Comparison between $\mathbf{x}_i$ and $\mathbf{s}_i$. The large red box (2) highlights improvement in capillary visibility. Small red box (1) points out a phantom vessel.



human $\mathbf{x}_i$ human $\mathbf{z}_i$ fish $\mathbf{x}_j$ fish $\mathbf{z}_j$

Figure 3.5: Examples of synthesized images. The latent image generated by the synthetic model considerably improves vessel appearance.

As a result, $\mathbf{s}_i$ is not appropriate for direct use in application, in spite of the desirable improvement they offer in visibility of capillaries (e.g., large red box). I use $\mathbf{s}_i$ as the target image in the synthesis model described in Section 3.3 while the original image $\mathbf{x}_i$ is the source image (i.e., $\mathbf{x}^s = \mathbf{x}_i$, $\mathbf{x}^t = \mathbf{s}_i$). Since the encoder f_e extracts causal features from $\mathbf{x}_i$, the noise in $\mathbf{x}_i$ as well as the phantom features in $\mathbf{s}_i$ will not exist in the latent image $\mathbf{z}_i$. Figure 3.5 displays examples of extracted latent images. It is visually evident that the latent image $\mathbf{z}$ successfully highlights the vasculature. Compared with the raw input, even delicate capillaries show improved homogeneity and separability from the background. Comparing with the original VICCE (Liu et al., 2020) presented in Section 3.3, the proposed algorithm does not require imaging on multiple different devices. Furthermore, the 3D extension of VICCE does not appear straightforward due to the differences in image spacing between OCT devices and the difficulty of volumetric registration in these very noisy images. In contrast, I propose to use self-fusion (Oguz et al., 2020) result on neighboring en-face images as the auxiliary image that has by construction perfect pixel-wise correspondence with the input required in the synthesis model. This removes the need for multiple devices or registration, and allows us to produce a 3D segmentation by operating on individual en-face OCT-A slices rather than a single depth-projection image.

3.4.1.3 Post-processing

To binarize the synthesized image $\mathbf{s}_i$ estimated by the encoder f_e, I apply the 2^{nd} Perona-Malik diffusion equation (Perona and Malik, 1990) followed by the global Otsu threshold (Otsu, 1979). Any islands smaller than 30 voxels are removed.

3.4.2 Experiments

3.4.2.1 Competing methods

Due to the lack of labeled data, no supervised learning method is applicable. Similar to my approach that follows the enhance + binarize pattern, I apply Frangi's multi-scale vesselness filter (Frangi et al., 1998) and optimally oriented flux (OOF) (Zhang et al., 2019; Law and Chung, 2008) respectively to enhance the artifact-removed original image, then use the same binarization steps described above. I also present results using Otsu thresholding and k-means clustering.

3.4.2.2 Implementation details

All networks are trained on an NVIDIA RTX 2080TI 11GB GPU for 50 epochs with batch size set to 2. For the first 3 epochs, the entire network uses the same Adam optimizer with learning rate of 0.001. After that, the encoder and decoder are separately optimized with starting learning rates of 0.002 and 0.0001 respectively in order to distribute more workload on the encoder f_e. Both networks decay every 3 epochs with at a rate of 0.5. I use the loss function presented in Equation 3.1, with hyper-parameters $\alpha = 1$ and $\beta = 0.05$.

3.4.3 Results

3.4.3.1 Qualitative results

The top row of Figure 3.6 illustrates 2D segmentation results within the manually segmented ROI, where LIFE can be seen to have better sensitivity and connectivity than the baseline methods. The middle and bottom row of Figure 3.6 show a 3D rendering of each method. In the middle row, I filtered out the false positives (FP) to highlight the false negatives (FN). These omitted FP areas are highlighted in yellow in the bottom row. It is easy to see that these FPs are often distributed along horizontal lines, caused by unresolved motion artifacts. Hessian-based methods appear especially sensitive to motion artifacts and noise; hence the Frangi filter and OOF introduce excessive FP. Clearly, LIFE achieves the best preservation in small capillaries, such as the areas highlighted in white boxes, without introducing too many FPs. Similar analysis is also conducted on the zebrafish data. Figure 3.7 indicates that LIFE has superior performance in capturing all small branches highlighted by the white boxes. Figure 3.8 shows the vessel tracking results from the vessel enhanced image using the segmentation result as seed, as well as from the original image using the K-

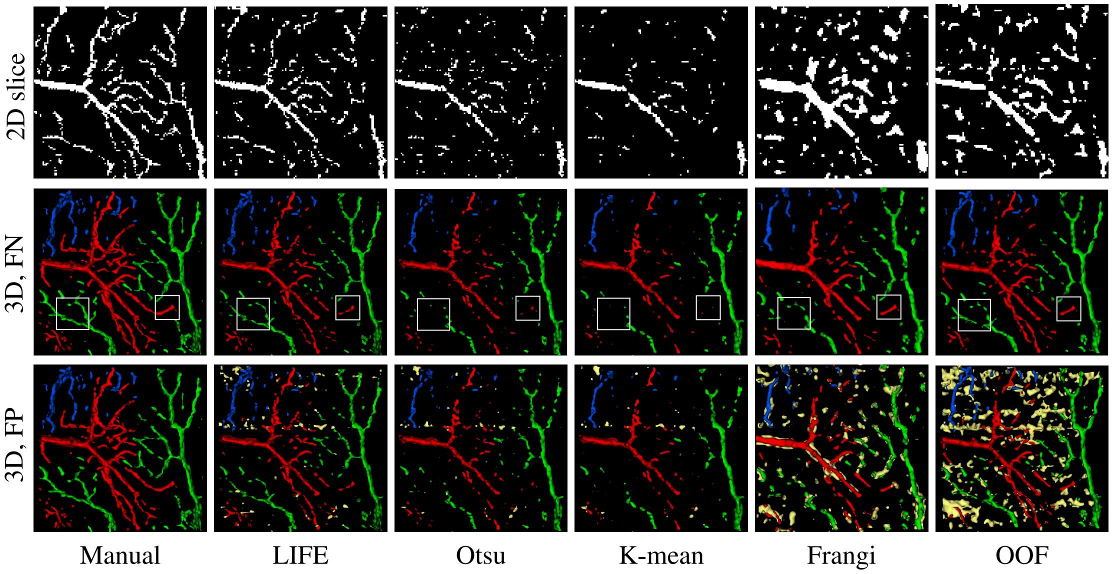


Figure 3.6: Human retinal OCT-A segmentation results. 2D slice and 3D marching cubes rendering of segmentation results on human retina, with Gaussian smoothing, $\sigma = 0.70$. Red, green, blue show three different branches; yellow highlights false positives. The proposed LIFE is the only method that can recover the 3D structure and connectivity of the capillaries outside the largest vessels without causing excessive FP (yellow). White boxes highlight LIFE's improved sensitivity.

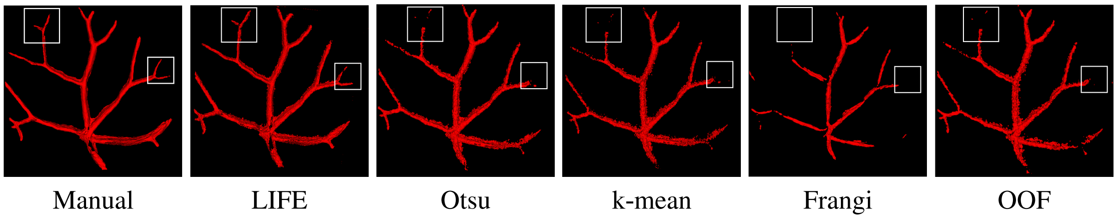


Figure 3.7: Segmentation results of zebrafish retina with the same rendering settings as in Fig. 3.6.

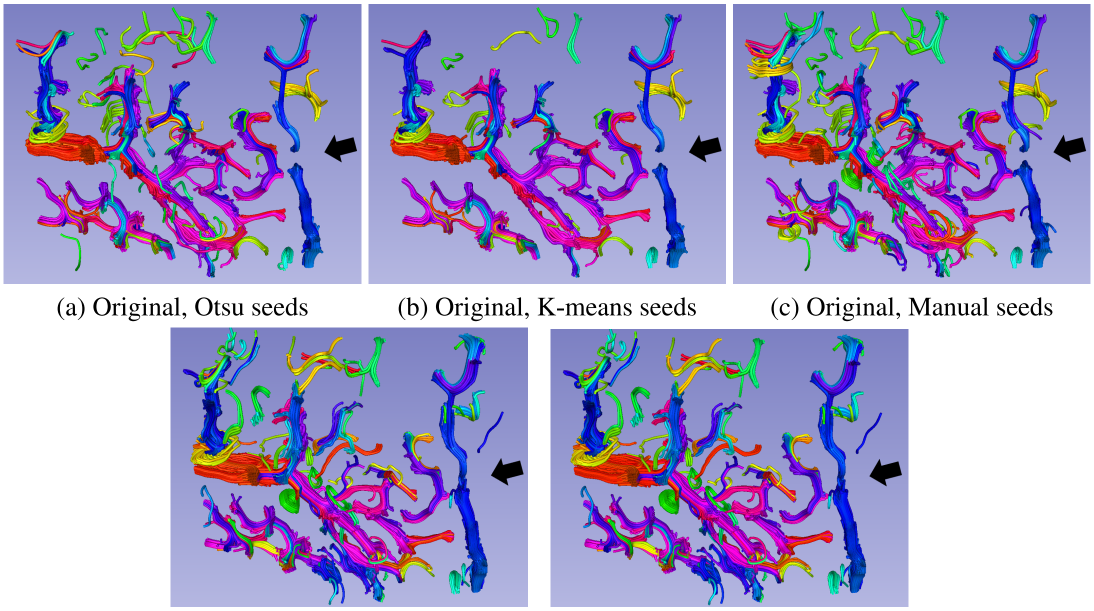
means, Otsu as seeds. The LIFE results are practically indistinguishable from the 'ideal truth' which uses the denoised LIFE image to compute the tensor field and the manual segmentation as labels. All methods using the original image to compute the tensor field suffer from connectivity problems.

3.4.3.2 Quantitative results

Figure 3.9 shows quantitative evaluation across B-scans, and Table 3.2 across the whole volume. Consistent with the qualitative assessments, LIFE significantly ($p \ll 0.05$) and dramatically (over 0.20 Dice gain) outperforms the baseline methods on both human and fish data.

3.4.4 Discussion

In this work, I propose a method that provides for the first time a 3D segmentation of fovea vessels and capillaries from OCT-A volumes. The introduction of the auxiliary image simulated by the self-fusion brings



(a) Original, Otsu seeds (b) Original, K-means seeds (c) Original, Manual seeds

(d) LIFE, Manual seeds (e) LIFE, segmentation seeds (mine)

Figure 3.8: Vessel tracking. Color indicates average vessel orientation (red: left-right, blue: up-down, green: out of plane). The image used to compute the Hessian field and the binary segmentation used as tractography seeds are indicated below each panel. The Otsu and K-means have poor connectivity and suffer from false negatives (a,b). LIFE can be seen to have both extensive capillary sensitivity and better connectivity than the baselines. It is especially noteworthy that the large vessel to the right **(black arrows)** is disconnected for all methods using the original tensor field, but connected for LIFE tensor fields. LIFE (e) is practically indistinguishable from the 'ideal truth' given by the LIFE tensor field and seeding with manual labels (d).

many benefits for the method. First of all, since this image is directly computed from the input data, no inter-volume registration is needed between the two modalities input to LIFE. Further, rather than purely depending on image intensity, it exploits local structural information to enhance small features like capillaries. Since LIFE works in a self-supervised manner, neither manual annotation nor multiple image acquisitions are required to train it. This enables my method to be easily applied to vessel segmentation on other modalities like magnetic resonance angiography. Still, there are some disadvantages to overcome in future research. For instance, LIFE cannot directly provide a binarized output and hence the crude thresholding method used in post-processing influences the segmentation performance.

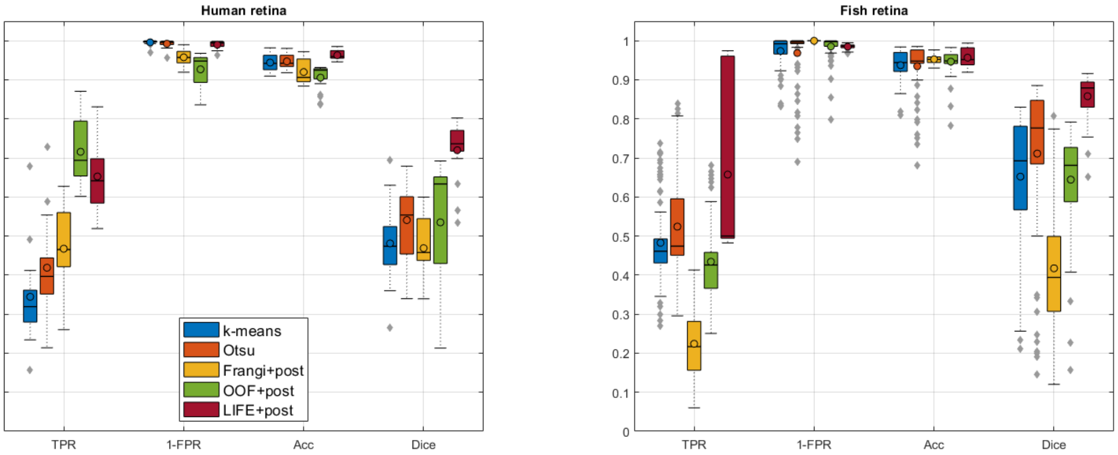


Figure 3.9: Quantitative result evaluation for (left) human and (right) zebrafish data. Evaluation is performed across B-scans. TPR: true positive rate, FPR: false positive rate, Acc: accuracy. LIFE achieves the best Dice performance for both datasets ($p \ll 0.05$).

Algorithm	TPR		FPR		Accuracy		Dice	
	Human	Fish	Human	Fish	Human	Fish	Human	Fish
k-means	36.33	41.67	**0.42**	3.03	94.40	92.28	51.52	62.49
Otsu	44.03	43.56	0.76	3.46	94.72	91.91	57.72	63.99
Frangi+bin	49.00	21.17	4.19	**0.02**	91.98	**94.89**	50.02	42.12
OOF+bin	**68.26**	37.75	7.48	1.65	90.53	93.50	54.14	62.47
LIFE+bin	66.13	**49.99**	1.04	1.53	**96.27**	93.86	**77.36**	**85.94**

Table 3.2: Quantitative evaluation of human and zebrafish segmentation (%). Evaluation is performed across the whole volume. TPR: true positive rate, FPR: false positive rate. Bold indicates the best score per column. LIFE achieves the best Dice performance for both datasets ($p \ll 0.05$).

3.5 2D Pseudo-modality Generation

3.5.1 Method

In this section, I present the application of the synthesis model on 2D fundus photography. As discussed in Section 3.3, the intuition of this approach is to extract shared features between two types of images with pixel-wise correlation. Given the annotated fundus photography images $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$, I treat the binary vessel mask $\mathbf{y}_i$ as the target image and the color fundus $\mathbf{x}_i$ as the source image. Unlike the previous scenario, the loss function used for training is no longer the image reconstruction loss. Instead, I use the segmentation loss composed by the cross-entropy loss and the Dice loss, defined as:

$$\mathscr{L}_{seg} = \underbrace{-\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{y}(i,j)\log\tilde{\mathbf{y}}(i,j)}_{\text{cross entropy loss}} + \underbrace{\left(1-\frac{2\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{y}(i,j)\tilde{\mathbf{y}}(i,j)}{\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{y}(i,j)^2+\tilde{\mathbf{y}}(i,j)^2}\right)}_{\text{Dice loss}} \tag{3.2}$$

where $\tilde{\mathbf{y}}$ is the binary prediction of the model, and (i,j) indicates the coordinate of pixels, H, W are the height and width of the image. Comparing with the previous case, the segmentation loss has a stronger ability to enhance the vessel in the synthesized image. Compared with OCT-A, the fundus photography contains many other anatomical structures such as optic disk, macular and lesions (e.g., hemorrhages and exudates). Given that the vessels are the only feature in the target image $\mathbf{y}$, these irrelevant structures can be suppressed or even removed by the encoder f_e.

Note that there is no direct constraint imposed on the latent representation. Therefore, the appearance/style of $\mathbf{z}$ can be pretty random while its anatomy is fixed, which is desirable for data augmentation. The randomness is purely introduced by the stochastic gradient decent (SGD) in the backward propagation. Hence, I take advantage of this degree of freedom by training 3 different models to generate three pseudo-modalities shown in Figure 3.10. As expected, most irrelevant features are removed and the vessels get enhanced. Given that these pseudo-modalities are identical in vessel structure while drastically diverse in

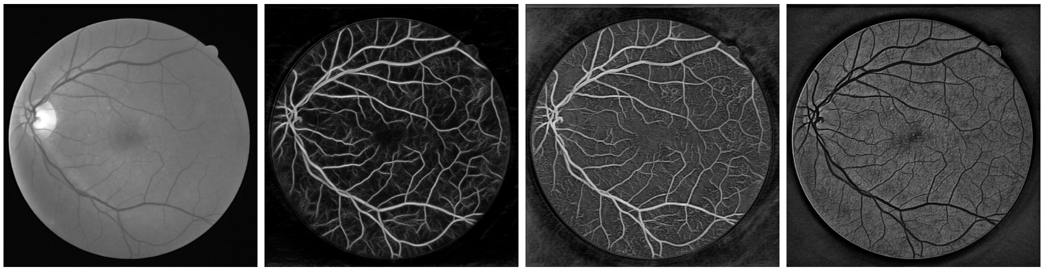

(a) $\mathbf{x}_i$ (b) $\mathbf{z}_i^1$ (c) $\mathbf{z}_i^2$ (d) $\mathbf{z}_i^3$

Figure 3.10: From left to right: (a) the green channel of an original color fundus image. (b-d) the synthesized latent images.

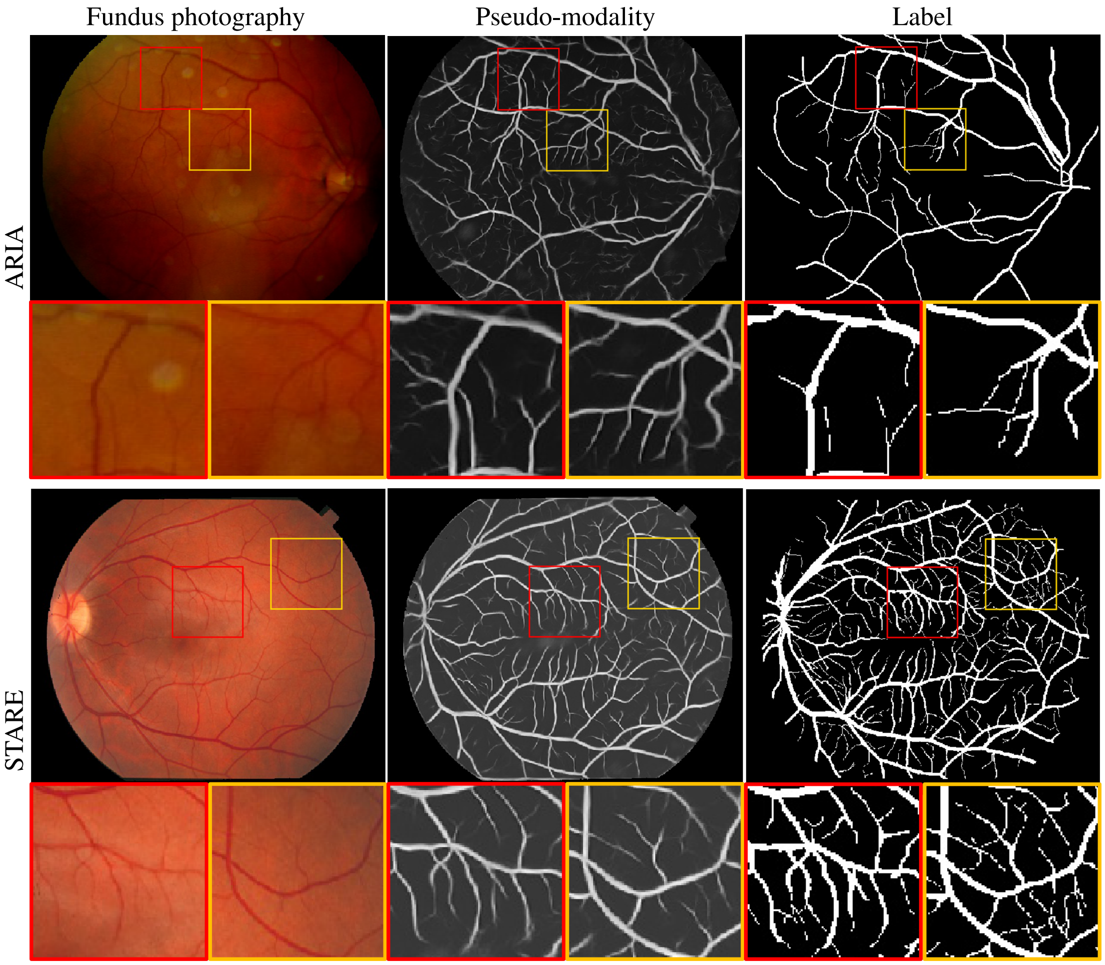


Figure 3.11: Test examples on two fundus photography datasets, ARIA (Farnell et al., 2008) and STARE (Hoover et al., 2000). Below each image, close-up panels show two highlighted areas (red and yellow boxes) for easier comparison. The proposed pseudo-modality provide excellent vessel clarity.

other aspects, they serve as valuable intermediate results well-suited for the domain generalization task that will be discussed in the next chapter.

3.5.2 Experiments

The network is trained and tested on an NVIDIA RTX 2080TI 11GB GPU. I use a batch size of 6 and train for 100 epochs. Patch-training is employed, I sample patches with size 256 × 256 as the input to the network. I use the Adam optimizer with the initial learning rate of 5×10^{-4}. The learning rate decays by 0.5 every 2 epochs. I empirically observed that a large learning rate can potentially result in black latent image.

3.5.3 Results

Figure 3.11 shows a example of the first pseudo-modality ($\mathbf{z}^1$) from two fundus photography datasets. I observe that for different datasets, the manual annotations includes varying amounts of detail: the label for the STARE dataset contains many more small vessels than ARIA. In the ARIA example, the pseudo-modality is able to enhance the thin vessels with very poor contrast. This is also evident by the big vessels seen at the bottom left quadrant of the image where the illumination is low. Moreover, the it filters out the circular artifacts seen within the red box. In the STARE example, the model extracts most of the vasculature including the faintly visible fine vessels. These tiny vessels have relatively lower intensity in the synthesized image, which suggests lower confidence. Compared with the manual labels, the pseudo-modality can better capture the vessels in some poor contrast conditions.

3.5.4 Discussion

In this section, I propose a synthetic model that can effectively extract a specific type of feature from the complex context. By utilizing the randomness of the stochastic gradient decent during the backward propagation, I am able to generate three stable types of images with well enhanced vasculature. A good property of these images is that they share the same vessel structure while the image styles are significantly different. Therefore, I named them as the anatomy-consistent pseudo-modalities and they can be implemented as augmented data in the domain generalization project that will be explored in Chapter 4.

3.6 Chapter Summary

In this section, I present a novel synthesis model to generate vessel enhanced 2D images given paired data with pixel-wise correlation. With the self-fusion method (Chapter 2), a self-supervised 3D vessel segmentation approach LIFE is developed. I show that it can significantly improve the segmentation performance compared to other unsupervised algorithms on both human and zebrafish retina OCT-A. However, due to the lack of 3D annotation, this evaluation is currently valid only on my in-house small datasets. This work was published in *MICCAI 2021* (Hu et al., 2021a) and the code is publicly available at https://github.com/DeweiHu/LIFE. Additionally, given labeled fundus photography datasets, I am able to generate three stable pseudo-modalities that share the identical vasculature while the image styles are drastically diverse. These results will prove to be beneficial for the domain generalization task in the following chapter. The code and the check points for the three synthesis models are available at https://github.com/MedICL-VU/Vector-Field-Transformer.

CHAPTER 4

Domain Generalization and Domain Adaptation

4.1 Introduction

Deep learning approaches have shown extraordinary potential in delineating patterns in high-dimensional data. However, the deep models are not guaranteed to work well on out-of-distribution (OOD) data, which soundly confines their applications. This problem attracts more attention in medical image analysis since even the data of the same modality acquired from multiple sites may distribute diversely caused by different scanners and imaging protocols. Such variation can result in distinct contrast and resolution for the output images. Moreover, the discrepancy in imaging modality has even more impact on data distribution. In Figure 4.1, I show examples illustrating such domain shifts. All these factors (the variation in image contrast, the unseen pathological phenotype, the change in image resolution and different image modality) severely degrade the performance of deep models trained for downstream tasks such as semantic segmentation and disease diagnosis. Therefore, either a domain adaptation (DA) (Guan and Liu, 2021) or a domain generalization (DG) (Zhou et al., 2022) method is often needed to enhance the robustness of the deep learning model on unseen data distributions.

For domain adaptation, the target domain data is accessible during the training process. This allows techniques such as fine-tuning on small labeled dataset from the target domain, or employing unsupervised methods when labels are not available, to directly adapt the model to the target domain. Among all the categories of DA, the unsupervised domain adaptation (UDA) is most commonly used as there is often no ground truth available for the unseen target data. Specifically for the segmentation task, manual annotation on medical images is particularly labor-intensive and time-consuming (Kumari and Singh, 2023). The UDA

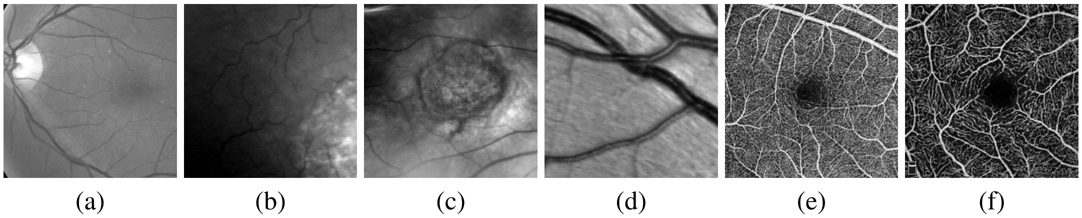

(a) (b) (c) (d) (e) (f)

Figure 4.1: Visual examples for domain shift on retinal vessel segmentation. From left to right, panels (a-d) are the green channel of fundus images (Staal et al., 2004; Farnell et al., 2008; Budai et al., 2013), panels (e-f) are OCT-A images (Li et al., 2020b; Ma et al., 2020). Suppose (a) represents the source domain. The distribution shift in the test domain can present as (b) poor contrast, (c) lesion, (d) resolution change, and (e, f) different modality.

can be particularly challenging when the domain shift is substantial (e.g., in different modalities). Another drawback for most of DA approaches is, for any unseen target dataset, an extra training step is often required. To reduce the processing time, the test-time adaptation approach (Liang et al., 2023) is investigated in recent research.

Unlike domain adaptation, domain generalization is defined under a more strict condition that the data from the target domain is entirely inaccessible during training. This turns out to be more realistic for medical images since a researcher is unlikely to have any sample data from other clinical centers while developing an initial model. Hence, some methods have been presented to solve the DG problem for medical image analysis, and most of these fall into three major categories. Firstly, there are data augmentation/generation based methods (Zhang et al., 2020b; Lyu et al., 2022). The training domain is expanded by applying handcrafted perturbations to the training data or by leveraging adversarial models to generate new data that is out of the current domain distribution. The second approach is domain alignment, which can be done on either image or feature space. The image space alignment refers to harmonization, which is usually achieved by image-to-image translation (Zuo et al., 2021). In the latter type of approach, additional constraints, such as KL divergence (Li et al., 2020a) and adversarial regularization (Aslani et al., 2020), are imposed to align the feature space. The last type of method is meta-learning, which is a general training strategy. After its introduction by Finn et al. (2017), the application of the episodic training paradigm has been extended to medical image analysis (Dou et al., 2019; Khandelwal and Yushkevich, 2020). In general, it mimics the condition when the model confronts data from a distribution unseen during training by dividing the source domains into meta-train and meta-test subsets. In this study, I investigate a workflow that marries the first two approaches to achieve DG for retinal vessel segmentation.

Although the retinal vessels are visualized with various appearances by different modalities such as fundus, OCT angiography, and fluorescein angiography, the tubular shape of the vessels remains a domain-agnostic feature that makes them recognizable for humans. The tubular shape, termed vesselness, has been mathematically modeled by a Hessian-based expression in (Frangi et al., 1998). Even though the learning-based models have outperformed the traditional Frangi filter in many aspects, the vesselness feature remains relevant since it can describe the essential character of vessels regardless of data distribution. Bridging the conventional handcrafted approaches, such as the Frangi filter, and the completely data-driven deep learning algorithms can be a good solution to the domain generalization problem. From a higher level standpoint, for both human and deep models, having some well-established prior knowledge involved in the training can be a better solution than learning from scratch.

Therefore, I propose to leverage a handcrafted feature map inspired by the Hessian description of the tubular morphology so that the model can discern the vessels by recognizing their shape in addition to intensities. Unlike the scalar vesselness feature computed from the eigenvalues in the Frangi filter, I implement the secondary eigenvector of the Hessian at each pixel as the geometric feature. In this way, I transform the original intensity image into a vector field. Ideally, the vectors within the vessel will be homogeneously oriented along the vessel direction, mimicking the blood flow. While computing the Hessian, I optimize the standard deviation of the Gaussian filter by maximizing the vesselness value presented by Frangi et al. (1998) to adapt the Hessian to vessels of various thicknesses. I regard this vector field as a common domain for data in different distributions. Such a vectorized feature is particularly suitable for the transformer model based on the cosine similarity attention mechanism. Hence, I introduce a specific model architecture called vector field transformer (VFT) in Section 4.3 to effectively leverage the correlation between eigenvectors in different ranges of context for vessel segmentation in 2D images. Then, inspired by the principles of the diffusion tensor imaging (DTI) (Le Bihan et al., 2001), I extend the VFT to a bipolar tensor field (BTF) in Section 4.4 to explicitly represent the vessel shape by a tensor at each pixel. Since both VFT and BTF are tubular shape representation based on the intensity gradient, they require the computation of Hessian matrix at each pixel which is time-consuming as the standard deviation of each Gaussian kernel is acquired by grid search. Furthermore, the Hessian matrix is inherently vulnerable to noise, which is the primary obstacle that prevents these methods from being conducted on low-quality data with poor contrast and/or high noise. To tackle these problems, I propose an implicit way of exploiting the morphological features by adopting the **m**eta-learning paradigm on **a**natomy-consistent **p**seudo-modalities (MAP) that are previously introduced in Chapter 3. This work is detailly presented in Section 4.5.

Other than the DG methods, I also present an unsupervised domain adaptation approach in Section 4.6 that enables the vessel segmentation model trained on one image modality to work on unlabeled target modalities. The general idea is to train a conditional diffusion probabilistic model (Ho et al., 2020) to synthesize target domain image from a given binary vessel mask. Similar models based upon a spatially-adaptive normalization block (Park et al., 2019) have been presented in data augmentation for histology (Yu et al., 2023; Oh and Jeong, 2023) and colonoscopy (Du et al., 2023). However, all these diffusion models are trained with annotated data in the same modality. In this study, I explore, for the first time, training the conditional diffusion model in a weakly supervised manner for the cross-modality scenario.

In this chapter, I first discuss literature of the domain generalization and domain adaptation algorithms. Then I describe the three DG algorithms and a DA approach in the following sections.

4.2 Related Work

4.2.1 Domain Generalization

4.2.1.1 Data augmentation

It is a common practice to alleviate the overfitting problem of complex deep models by introducing some perturbations to the homogeneously distributed training dataset. If such disturbances are added directly to the image, I name it *image transformation-based augmentation*. This approach has been thoroughly discussed in the computer vision literature (Volpi and Murino, 2019), and its application has proven helpful for the generalization of models trained on medical images as well (Otálora et al., 2019; Chen et al., 2020a). Zhang et al. (2020b) leverage the deep stacked transformations named the *BigAug*, which includes a series of augmentation methods to introduce variation in the quality, appearance, and spatial configuration of the images. They evaluate the method by the downstream 3D segmentation performance on MRI and ultrasound images. In (Lyu et al., 2022), the authors further incorporate data augmentation in the optimization process of the model by reinforcement learning.

Other than image transformation, there also is *model-based augmentation* that transfers the style of images by a neural network. Xu et al. (2020b) use a multi-scale random convolution layer to generate a feature map with a random style. The mixture of the feature map and the input image is regarded as a pseudo-novel domain with the original semantics preserved. In my work, I utilized the anatomy-consistent pseudo-modalities presented in Chapter 3 as the synthesized augmentation data following diverse distributions.

4.2.1.2 Domain alignment

Another common strategy for DG is to align the distribution of unseen data with the source domain so that the model can provide a robust prediction. This approach can be conducted on both image and feature space. The former scenario refers to image harmonization, which is usually achieved by image-to-image translation. With data from one site specified as the reference, the distribution shift is alleviated by mapping data from other sites to the reference domain. MRI is the most common modality for harmonization applications since it has more flexibility in terms of acquisition parameters, which induces heterogeneous image contrast (Dewey et al., 2020; Zuo et al., 2021). Although harmonization is widely utilized on various modalities (e.g., CT (Selim et al., 2021) and OCT (Ren et al., 2021)), most proposed methods require access to data from all sites in the training phase. Therefore, harmonization does not allow adaptability to unseen domains. Thus it can be sensitive to domain shifts despite being designed to be a solution to it (Zuo et al., 2021).

The latter case is to unify the feature space of OOD data with that of source domain data. This is usually accomplished by an adversarial network. Aslani et al. (Aslani et al., 2020) indicate that the latent features of inputs from different distributions are inherently disentangled from each other and can be classified by a regularizer. Thus, they introduce three loss functions to re-entangle the latent features so that the regularizer can not differentiate them. In my work, I propose to leverage a Hessian-based vector field to align the distribution in terms of tubular shape description. This vector field can be regarded as a handcrafted feature space inspired by the Frangi filter (Frangi et al., 1998).

4.2.1.3 Meta learning

Meta learning is a booming area in recent research for DG (Shu et al., 2021; Zhou et al., 2022; Qin et al., 2023), inspired by the episodic training paradigm introduced by MAML (Finn et al., 2017). Usually, the training data is divided into meta-train and meta-test sets to mimic the condition when the model confronts an unseen dataset. The goal is to improve the performance on the meta-test sets. Khandelwal and Yushkevich (2020) apply the basic episodic training paradigm on the segmentation of different parts of vertebrae. Dou et al. (2019) introduce an additional global loss function to delineate the inherent consistency between meta-train and meta-test sets for the multi-class classification problem. Their approach is further extended to tissue segmentation in multi-site brain MRI.

4.2.2 Unsupervised Domain Adaptation

4.2.2.1 Feature alignment

The UDA can be achieved by reducing the disparity between intermediate feature maps extracted from source and target domain images. Such features are also regarded as a domain-invariant representation. In practice, this idea primarily relies on the mini-max approach in adversarial learning (Goodfellow et al., 2014). In the domain-adversarial neural network (DANN) (Ganin and Lempitsky, 2015), a gradient reversal layer is implemented in an adversarial network framework to enforce the comparability of feature map distributions across domains. Based on DANN, Javanmardi and Tasdizen (2018) propose a domain adaptation method on fundus photography vessel segmentation.

4.2.2.2 Image alignment

In contrast, image alignment methods perform domain alignment in image space, by converting an image in source domain to the style of the target domain. Here, 'style' refers to the appearance characteristics excluding the semantic content. Such unpaired image-to-image style transfer is commonly done using CycleGAN (Zhu et al., 2017). Palladino et al. (2020) implement this framework in white matter segmentation for multicenter MR images. Huo et al. (2018) propose an end-to-end framework called SynSeg that combines the cycle adversarial network with the segmentation model.

4.3 Explicit shape modelling: Vector field transformer

In this section, I will describe a method to model the morphology of the vessels with a Hessian-based vector field which is regarded as the common domain for data in different distribution. Based on this shape representation, I propose a vector field transformer to let the model learn from the correlation between the vectors in a small neighborhood. To improve the robustness of the model, I apply the anatomy-consistent pseudo-modalities presented in Section 3.5.

Figure 4.2 shows the overall pipeline of the proposed method. There are three major steps. First, I implement the pseudo-modalities introduced in Section 3.5 as the augmented data to expand the source domain (FP green channel images). Next, I connect the deep learning approach with the classic model-based algorithm by constructing a vector field to represent the shape of retinal vessels. Finally, I present a new transformer architecture that can take advantage of the shape information to segment the vessels without being severely affected by domain shifts.

4.3.1 Methods

4.3.1.1 Hessian-based vector field

Although the traditional model-based algorithms are surpassed in performance by deep learning methods, their generalizability and interpretability remain highly relevant as they mathematically model the basic human intuition about blood vessels. Therefore, I aim to find a way to fuse the traditional algorithms with the data-driven approach to solve the DG problem. Unlike other segmentation tasks (e.g., subcortical structures, multiple sclerosis lesions, etc.) which can have wide range of geometric properties, the shape of vessel segments is always tubular and can thus be easily modeled. The only factor that varies for different segments (away from branching points) is the diameter of the tubular shape. Thus, the Frangi filter (Frangi et al., 1998) and its variations (Canero and Radeva, 2003; Jerman et al., 2015) provide a multiscale measurement of the vesselness for both 2D and 3D scenarios. These algorithms use the eigenvalues of the Hessian matrix at

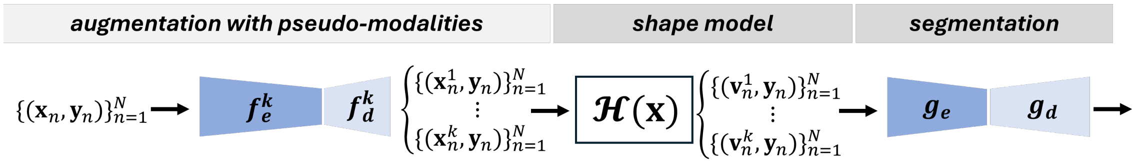

augmentation with pseudo-modalities
shape model
segmentation
$\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$
f_e^k
f_d^k
$\{(\mathbf{x}_n^1, \mathbf{y}_n)\}_{n=1}^N$
$\vdots$
$\{(\mathbf{x}_n^k, \mathbf{y}_n)\}_{n=1}^N$
$\mathcal{H}(\mathbf{x})$
$\{(\mathbf{v}_n^1, \mathbf{y}_n)\}_{n=1}^N$
$\vdots$
$\{(\mathbf{v}_n^k, \mathbf{y}_n)\}_{n=1}^N$
g_e
g_d

Figure 4.2: The overall pipeline of an the proposed algorithm. I augment the data by synthesizing anatomy-consistent pseudo-modalities introduced in Section 3.5. Then, the vessel shape modeling is achieved via vector field construction. Finally, the segmentation network is a vector field transformer. $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ represent the paired data where the subscript n indicates the image index and N is the total number of images in the dataset. The superscript k in $\mathbf{x}_n^k$ denotes the index of pseudo-modality. In my experiment, $k = 3$. $\mathbf{v}_n^k$ represent the vector field computed from $\mathbf{x}_n^k$.

each pixel to depict the homogeneous intensity value along the vessel as well as the strong gradient in the orthogonal direction. In the 2D case, a larger ratio $|\lambda_2|/|\lambda_1|$ usually indicates higher vesselness. Note that in this work I denote the eigenvalues in ascending order in terms of absolute value i.e., $|\lambda_1| \leq |\lambda_2|$.

To combine this classic shape model with deep learning, I introduce a Hessian-based vector field which takes advantage of both eigenvalues and eigenvectors of the Hessian and create an implicit description of tubular structure. Following the same intuition as the Frangi filter (Frangi et al., 1998), I observe that the minor eigenvectors of the Hessian $\mathscr{H}$ form a smooth vector field with streamlines that follow along the retinal vessels mimicking the blood flow. Ideally, the vectors within the vessel should be parallel in orientation; however, those in the background are a lot more random. Suppose $\mathbf{x}(i,j)$ denotes the intensity of the pixel at coordinate (i,j) in a 2D image $\mathbf{x} \in \mathbb{R}^{H\times W}$. Considering the vessels vary in diameter, I optimize the Hessian at each pixel $\mathscr{H}(\mathbf{x}(i,j),\sigma^*)$ with regard to the 2D vesselness $\mathscr{V}(\sigma)$ within a range of scales:

$$\sigma^* = \underset{\sigma_{min}\leq\sigma\leq\sigma_{max}}{\operatorname{argmax}} \; \mathscr{V}(\sigma) \tag{4.1}$$

where the vesselness is defined in the same way as originally proposed in (Frangi et al., 1998):

$$\mathscr{V}(\sigma) = \begin{cases} 0 & \text{if } \lambda_2 > 0, \\ \exp\left(-\frac{R_B^2}{2\beta^2}\right)\left[1-\exp\left(-\frac{S^2}{2c^2}\right)\right] & \text{else} \end{cases} \tag{4.2}$$

Here $R_B = \lambda_1/\lambda_2$, S is the Frobenius norm of the Hessian ($\|\mathscr{H}\|_F$), $\beta = 0.5$ and $c = 0.5$. Note that I assume the vessels are bright and the background is dark in this work. For cases like the green channel of fundus photography image, I flip the intensity before normalizing the intensity value to range $[0,1]$. Then the optimized Hessian is represented by a 2×2 matrix:

$$\mathscr{H}(\mathbf{x}(i,j),\sigma^*) = (\sigma^*)^2\mathbf{x}(i,j) * \nabla^2 G(\mathbf{x}(i,j),\sigma^*) \tag{4.3}$$

where $G(\mathbf{x}(i,j),\sigma^*)$ is a 2D Gaussian kernel. Then I apply the eigen decomposition to obtain the minor eigenvector $\mathbf{v}^1(i,j)$ which is a unit vector. To combine the shape descriptor with the intensity information, I use the normalized intensity value as the magnitude of the vector, which yields the vector field $\mathbf{V}(i,j) = \mathbf{x}(i,j)\mathbf{v}^1(i,j)$. This vector field $\mathbf{V} \in \mathbb{R}^{2\times H\times W}$ is used as the input to the segmentation network.

One promising characteristic of this vector field representation is that it provides a consistent geometrical delineation for vessels across different modalities, as it emphasizes the structural information rather than pure image intensity. To show that this property holds across diverse domains, I plot the vector field of images from different datasets in Figure 4.3. There are two attributes of the vectors, magnitude and orientation.

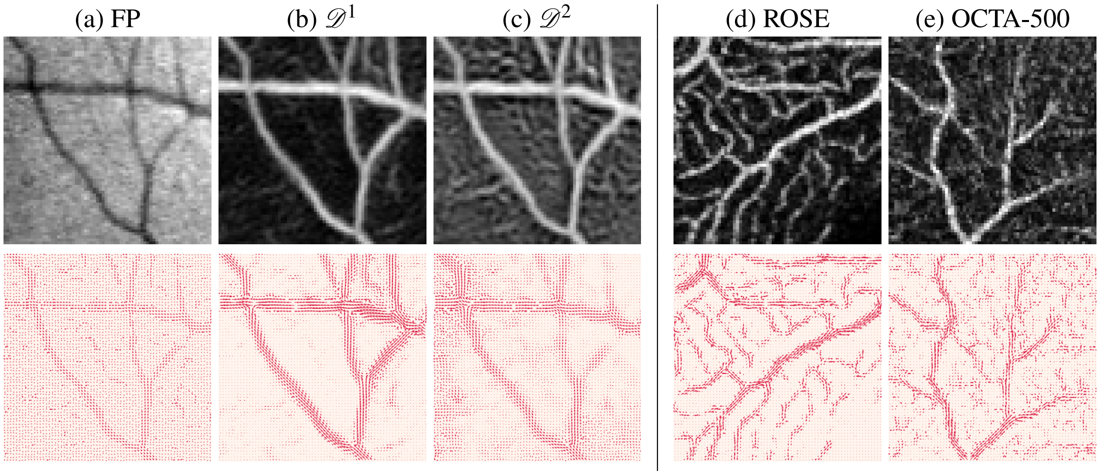


Figure 4.3: Vector field of different domains. The top row shows the example patches from 5 different domains. The bottom row are the corresponding vector fields. The vectors inside the vessels are coherent in magnitude and orientation across domains. In (a), the raw FP green channel image is cropped from an image in DRIVE (Staal et al., 2004). (b) and (c) are pseudomodalities derived from (a). Finally, (d) and (e) are from ROSE (Ma et al., 2020) and OCTA-500 (Li et al., 2020b), respectively.

The magnitude of the vector is determined by the image intensity and hence tends to vary across different domains. In other words, the magnitude is sensitive to domain shift. Here I want to stress on the vector orientation which is always aligned with the vessel direction. Therefore, I augment the data as described in Section 3.5, to generate images with the same vessel structure while significantly different in intensity distribution. In this chapter, I denote the three pseudo-modalities with $\mathscr{D}^1$, $\mathscr{D}^2$ and $\mathscr{D}^3$.

In the training phase, the model will focus on the vector orientation. As illustrated in Fig. 4.3 columns (a), (b), and (c), the vectors inside the vessels are coherent in magnitude and orientation regardless of the difference in image intensity values across domains. Obviously, the raw FP green channel image has poor contrast and strong noise in the background, which yields a weaker vector field compared with that of the OCT-A images in columns (d) and (e). Nevertheless, the vector field representations are shown to have significantly bridged the gap between the two modalities. To effectively utilize these feature vectors, I propose a transformer model that takes the vector field as input in the following section.

4.3.1.2 Vector field transformer

Originated from natural language processing (Vaswani et al., 2017), the transformer architecture has been vastly deployed on image analysis (Dosovitskiy et al., 2020) and has outperformed its CNN counterparts (Hatamizadeh et al., 2022; Li et al., 2022). The transformer blocks are able to depict the correlation between features extracted from far-away patches in terms of cosine similarity. Feature embedding is usually required to reduce the computational cost. In this work, I introduce a vector field transformer that only focuses on

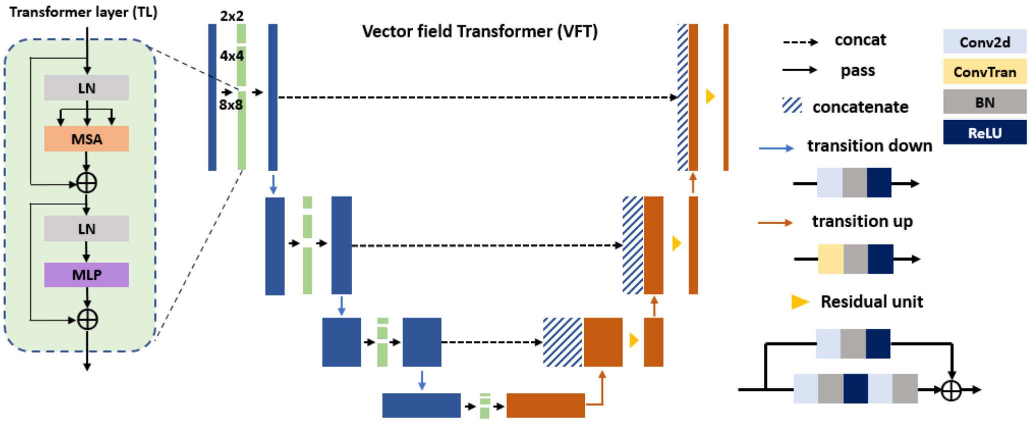


Figure 4.4: The network architecture of the VFT.

the context *within* the partitioned patches instead of the potential correlation *between* them. No feature embedding is required since the vectors have already represented the structural pattern.

As indicated in the Figure 4.4, the backbone of the neural network is a residual U-Net that takes as input a vector field with dimension $\mathbf{V} \in \mathbb{R}^{2\times256\times256}$. In the encoder, the residual unit in each layer is replaced by 3 paralleled transformer blocks which will not change the dimension of the input tensor (as detailed in the following paragraph). The downsampling is achieved by a transition-down-block that contains a 2D convolution layer with 4×4 kernel and stride of 2, a batch normalization layer, and an exponential linear unit (ELU). This transition-down-block will increase the channel number while reducing the height and width of the feature map by half. Denote the channel number n for each layer by C_n. In my experiments, I apply a 5-layer model with channel number $\{C_2 - C_{16} - C_{32} - C_{64} - C_{64}\}$ (i.e., the bottleneck feature map $\mathbf{z} \in \mathbb{R}^{64\times16\times16}$). Similarly, in the decoder g_d, I apply the transpose 2D convolution layer with 4×4 kernel and stride of 2, batch normalization, and ELU in the transition-up-block. Each layer in the decoder includes a residual unit.

I incorporate the transformer blocks (TB) in each layer of the residual U-Net encoder. In order to capture the vector orientation similarity in different scales of context, I break the image into three types of patches (2×2, 4×4, 8×8) as shown in Figure 4.5. Each type will go through an individual TB in a parallel fashion. In most existing vision transformer networks, e.g., (Dosovitskiy et al., 2020; Chen et al., 2021), it is common to embed the images to the feature space and apply the self-attention mechanism to find the correlation between different patches. In contrast, I discard the feature embedding step since the vector field can sufficiently represent the feature space. The cosine similarity in-built in the multi-head self-attention (MSA) can measure the alignment of vectors for vessel recognition. Moreover, unlike the semantic segmentation on natural images, the vessels are usually fine-scale features that require local attention. Thus, instead of comparing the

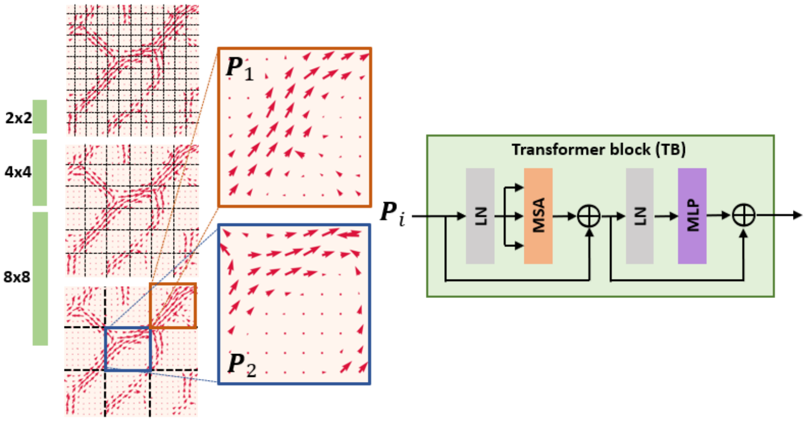


Figure 4.5: The paralleled transformer blocks with different window sizes. LN: layer normalization, MSA: multi-head self-attention, MLP: multi-layer perceptron. This image is an example for the very first layer of VFT.

similarity between patches, I implement the transformer block within each window. For example, in Fig. 4.5, P_1 and P_2 are two 8×8 patches in an image. The coherence of vectors within a patch (e.g., P_1) is more important for vessel recognition than the similarity between patches (e.g., P_1 and P_2). Finally, the paralleled multi-scale pathways allow the detection of vessels in diverse diameters.

Given the input $P_i \in \mathbb{R}^{C \times H \times W}$, the output of each TB has the same dimension (i.e., $\text{TB}(P_i) \in \mathbb{R}^{C \times H \times W}$) where C denotes the length of the feature vectors in each layer of the encoder. I concatenate all three outputs by channel and apply a 2D convolution to linearly map it back to $\mathbb{R}^{C \times H \times W}$. I leverage the same transformer block structure as proposed in (Chen et al., 2021). I set the number of heads in the MSA to $\frac{C}{2}$. The output of the multi-layer perceptron is set to be $4C$.

4.3.2 Datasets

To thoroughly evaluate the DG performance of the model, I consider several types of domain shift and test them accordingly using the publicly available datasets in Table 4.1. The color code in the table indicates the similarity between datasets. The first type of domain shift is the drastic change of resolution which can happen for multi-site data. This matters for the downstream segmentation task since the high-resolution images usually have vessels in substantially larger diameters that have never been seen in the source domain. The high-resolution fundus (HRF) image (Budai et al., 2013) is thus marked in light gray in Table 4.1 as it is significantly different than other fundus datasets in image size. The second scenario for domain shift is across imaging modalities. The DRIVE (Staal et al., 2004), STARE (Hoover et al., 2000), ARIA (Farnell et al., 2008) and HRF datasets are color fundus images while the OCTA-500 (Li et al., 2020b) and ROSE (Ma et al., 2020) are OCT angiography. The last case of domain shift is disease status. Note that the ARIA and

	modality	**resolution**	**# sample**
DRIVE	fundus	565×584	20
STARE	fundus	700×605	20
ARIA	fundus	768×576	$61/59/23$
HRF	fundus	3504×2336	$15/15/15$
OCTA-500(6M)	OCTA	400×400	300
ROSE	OCTA	304×304	30

Table 4.1: Datasets used. I use the color to code the intuitive similarity between dataset distributions. OCTA-500 Li et al. (2020b) contains two subsets: OCTA_6M and OCTA_3M. I use OCTA_6M, which includes samples with larger field of view (6mm×6mm×2mm).

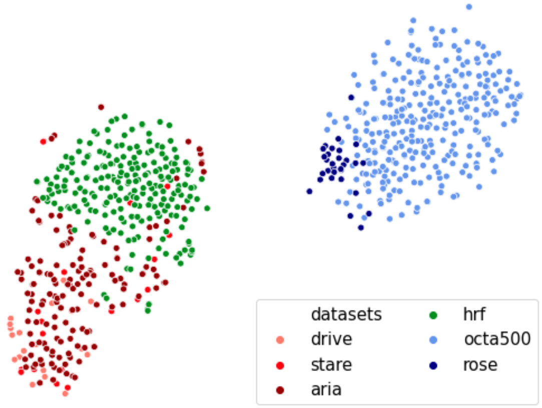


Figure 4.6: To visualize the distribution of raw data and the synthetic datasets, I first map the image to a feature vector $\mathbf{l} \in \mathbb{R}^{2048\times 1}$ by a pre-trained VGG16 to reduce dimension. Then I apply the t-SNE on the feature vectors. The three different types of red dots indicate the relatively low-resolution fundus datasets. The green dots represent the high-resolution fundus. The blue dots are OCT-A datasets. This t-SNE plot illustrates that these three conditions are generally separated in terms of distribution and the cross-modality scenario has larger displacement which is consistent with the intuition.

HRF have class labels for disease status, thus I list the number of samples for each class in Table 4.1. For ARIA, the classes are healthy/diabetic/AMD while for HRF they are healthy/diabetic/glaucoma.

Figure 4.6 illustrates the inherent variation in distribution of the raw data. Consistent with the color code in Table 4.1, the change in resolution induces a moderate amount of domain shift (light/dark red vs. green clusters) while the modality difference can cause a large gap between domains (blue/dark blue vs. others). I implement the pre-trained VGG-16 model (Simonyan and Zisserman, 2014) to extract feature vectors and apply the t-SNE (Van der Maaten and Hinton, 2008) for visualization.

4.3.3 Experiments

4.3.3.1 Implementation Details

All networks are trained and tested on an NVIDIA RTX 2080TI 11GB GPU. I use a batch size of 6 and train for 100 epochs. Patch-training is employed, I sample patches with size 256×256 as the input to the network.

	FLOP	#parameters
+Regularizer	5.42×10^9	760855
+BigAug	5.42×10^9	743150
+MASF	5.34×10^9	740614
Proposed	7.20×10^9	440854

Table 4.2: The number of floating point operations (FLOP) and the number of parameters for each model.

I use the Adam optimizer with the initial learning rate of 5×10^{-4}. The learning rate for decays by 0.5 for every 2 epochs.

4.3.3.2 Competing Methods

As discussed in Section 4.2, there are three major types of approaches for DG, and I pick one representative algorithm from each type as a competing method. Firstly, for data augmentation, I implement BigAug (Zhang et al., 2020b). Next, the MASF (Dou et al., 2019) is the one I select to represent the meta-learning solution for DG. Lastly, I use the domain regularization (Aslani et al., 2020) as the example of feature alignment via adversarial network. The baseline model is a vanilla residual U-Net trained only on the source domain $\mathscr{S}$ without attempting any DG, and the oracle model is trained directly on each target domain to represent the best achievable performance.

Note that I set the total number of parameters in the residual U-Net to be more than that of the VFT for fair comparison ($7.43 \times 10^5 : 4.40 \times 10^5$). In Table 4.2, I compare the complexity of models in by number of parameters and floating point operations (FLOP). The proposed method has fewer number of parameters, while there are more operations conducted to compute the similarity between feature vectors at each location. Note that VFT takes the vector field as input, hence the Hessian and eigenvectors need to be computed during pre-processing. Consequently, the VFT does not have advantage in inference speed.

4.3.3.3 Source and Target domains

To test the DG performance on different domain shift scenarios, I conduct experiments in two types of settings (Table 4.3).

For type I, all the source domains are color fundus images, and the target domains include all three cases of domain shift, i.e., differences in resolution, modality and disease status. Hence, I will be primarily focusing on the type I setting in the following sections. Type II experiments illustrate the performance of the model when trained purely on OCT angiography data and tested on fundus images. This setting is less prevalent in practice since there are more publicly available annotated fundus datasets, so in practice it is more common to train on fundus images instead of OCT-A.

	source domains ($\mathscr{S}$)	**target domains** ($\mathscr{T}$)

Type I	DRIVE(all)
	STARE(all)
	ARIA(healthy)
		ARIA(disease)

		HRF (all)
		ROSE OCTA-500(6M)

Type II	ROSE
	OCTA-500(6M)
		DRIVE(all)
		STARE(all)
		ARIA (all)

		HRF (all)

Table 4.3: Two types of experiments for training and testing datasets. 'All' refers to datasets where healthy subjects and patients are pooled together. I note that ROSE and OCTA-500 datasets do not specify disease status.

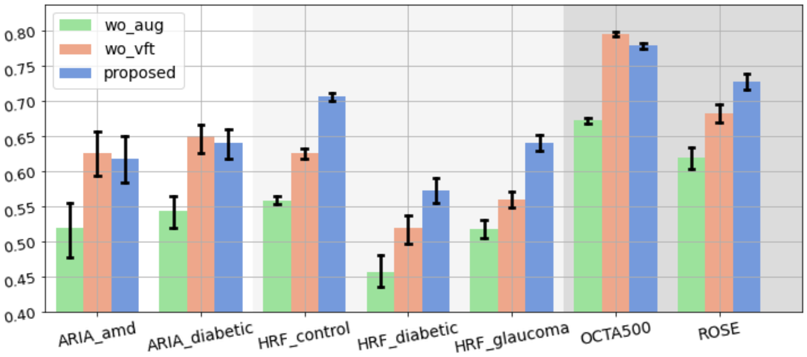


Figure 4.7: Ablation study results. The background is color-coded in the same way as Table 4.1. The blue bar is always higher than the green one which indicates that the augmentation significantly improves the performance. Comparing the red and blue, I can conclude that the proposed VFT architecture helps a lot for cross-resolution cases (HRF).

4.3.4 Results

4.3.4.1 Ablation Study

In the ablation study, I use the type I experimental setting to compare the proposed method with two other cases: (1) VFT model trained on raw data, without augmentation. (2) residual U-Net model trained on augmented vector field. The former scenario is used to prove the effectiveness of the synthetic data augmentation by the pseudo-modalities, while the latter condition is designed to show the usefulness of the novel transformer architecture. I note that the contribution of the vector field, the third main component of my approach, is discussed in the next section in the comparison against competing models.

The results for the ablation study are shown in Figure 4.7. The background of the datasets is color-coded in the same way as Table 4.1, i.e., white indicates shift in disease status, light gray indicates shift in resolution, and dark gray indicates shift in image modality. The green and red bars correspond to the aforementioned two ablation cases (VFT model trained on raw data, without augmentation, and residual U-Net model trained on augmented vector field, respectively), while the blue bars represent the proposed method in its entirety. Clearly, the data augmentation helps a lot in all the target domains. The proposed transformer architecture significantly improves the robustness of the model when confronted with OOD data with substantially increased resolution (light gray background) while its contribution is less clear in the other OOD cases.

4.3.4.2 Comparison with Competing Methods

I test the model in both type I (fundus to OCT-A) and type II (OCT-A to fundus) settings and compare with three existing solutions for DG as described in Section 4.3.3. I regard these methods as add-ons to the baseline model, so I implement the residual U-Net backbone for all these approaches.

Type I In Figure 4.8 I show the qualitative results in which the red and green represent the false negative (FN) and false positive (FP), respectively. The rows illustrate an example from each target domain, i.e., ARIA (diabetic), HRF (control), OCTA-500 and ROSE. Each column compares the result from different model. In the ARIA example, I show that the Hessian-based VFT is more sensitive to the insufficient image contrast, so it fails to capture the vessels in the shadow. The HRF dataset represents the cross-resolution scenario, so the example in the second row is in shape 2250pix × 2250pix. Most methods miss the major vessel trunk even though it is very prominent, since the source domain does not include vessels of similar thickness, given the resolution change of HRF. My approach performs better for such thick vessels with the help of the vector field. I note, however, that it is not as good as the oracle since the vessel diameter is out of the search range $[\sigma_{min}, \sigma_{max}]$ when computing the optimal Hessian. For the cross-modality case, VFT tends to provide a cleaner segmentation with less FP compared to the competing methods.

Tables 4.4, 4.5, 4.6 and 4.7 show the Dice coefficient, accuracy, recall and the normalized Hausdorff distance, respectively. The Dice scores indicate that the proposed method provides the best segmentation for HRF, OCTA-500 and ROSE datasets indicating its superior generalizability in cross-resolution and cross-modality conditions. These results are significantly better than that of the baseline (marked with †, with the significance p-value is computed by two-sample t-test). The improvement over the competing methods is especially pronounced in the HRF dataset, where the ablation study also showed the VFT is really able to shine (Figure 4.7). Table 4.5 shows that VFT has good accuracy across all target domains. On ROSE, it performs even better than the oracle model. Table 4.6 indicates that the my method has higher true positive rate for cross-site datasets with different image resolutions. The Hausdorff distance results in Table 4.7 show similar trends.

However, for the pathological samples in ARIA, the VFT performs worse than the baseline in both Dice score and accuracy, although this difference does not reach statistical significance ($p \geq 0.05$, marked with $\sim$). I observe that in the ARIA dataset, there are many images with poor contrast and illumination, such that

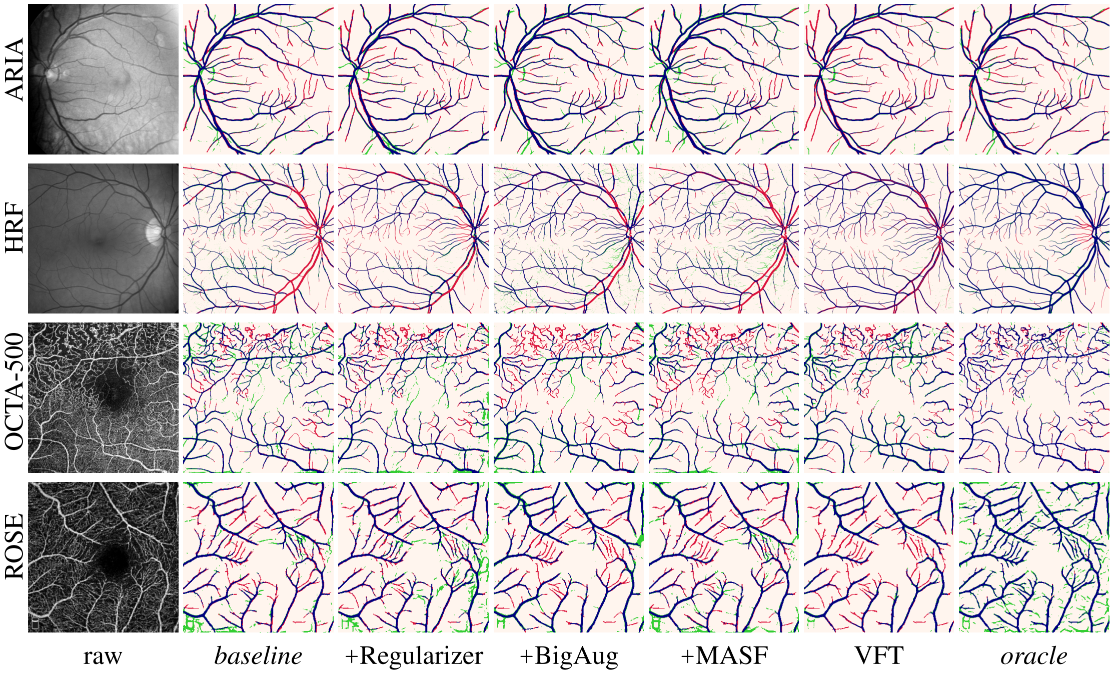


Figure 4.8: Qualitative results. Red and green represent false negatives and false positives, respectively. The navy blue is the true positive.

even the large vessels are indistinguishable from the dark background. This turns out to be a fatal problem for my approach since the Hessian-based vector field is no longer able to delineate the vessels due to the lack of contrast. Moreover, my synthetic augmentation does not provide any poor contrast samples. Consequently, the proposed method fails to get any improvement on the ARIA dataset compared to the baseline. Besides, it is easy to observe from the HRF results that the disease status (e.g., diabetic and glaucoma) can severely affect the vessel segmentation, which is likely another contributor to the poor performance in the ARIA test cases.

Type II In Tables 4.8, 4.9, 4.10 and 4.11, I similarly present the results for the OCTA-to-fundus setting. The presented method achieves the best in DRIVE, STARE and HRF datasets while it still struggles to significantly outperform the baseline in the ARIA dataset, consistent with the Type I test. Due to the poor contrast condition in ARIA dataset, the overall accuracy is lower, and none of the models consistently provide accurate prediction. Since the source domain only has OCT-A images, all the test samples are cross-modality. Therefore, I observe that the results for the same datasets (e.g., HRF) are not as good as they were in the Type I test. Training on OCTA and testing on fundus appears to be a more difficult task, perhaps because the fundus image contains many prominent features with relatively strong edges, such as the optic disc and lesions, in addition to vessels. These structures can cause confusion for any models that have not seen such features in training. Thus, the cross-model Dice score in Type II setting is generally lower than Type I for all

Method	ARIA		HRF			OCTA-500	ROSE
	amd	diabetic	healthy	diabetic	glaucoma		
baseline	0.6598	0.6815	0.6406	0.5267	0.5566	0.7316	0.6741
+Regularizer	0.6489	0.6697	0.6403	0.5216	0.5625	0.7354	0.6836
+BigAug	**0.6741**	**0.6932**	0.6613	0.5389	0.5735	0.7688	0.6932
+MASF	0.6709	0.6899	0.6598	0.5297	0.5720	0.7507	0.6727
Proposed	0.6181$^{\sim}$	0.6405$^{\sim}$	**0.7058**†	**0.5732**†	**0.6410**†	**0.7791**†	**0.7281**†
oracle	0.7334	0.7065	0.8358	0.7524	0.7732	0.8657	0.7603

Table 4.4: The Dice values for the target domains of type I test. The boldface indicates the best performance and the underline marks the second best result. $\sim$: p-value $\geq$ 0.05, $\dagger$: p-value $\ll$ 0.05 with regard to the baseline output. The background is color-coded in the same way as Table 4.1.

Method	ARIA		HRF			OCTA-500	ROSE
	amd	diabetic	healthy	diabetic	glaucoma		
baseline	0.9522	0.9536	0.9441	0.9389	0.9440	0.9414	0.9079
+Regularizer	0.9506	0.9516	0.9450	0.9408	0.9460	0.9469	0.9074
+BigAug	0.9470	0.9472	0.9406	0.9310	0.9375	0.9529	0.9138
+MASF	**0.9543**	**0.9561**	0.9452	**0.9471**	**0.9510**	**0.9603**	0.9164
Proposed	0.9407†	0.9521†	**0.9483**†	0.9384$^{\sim}$	0.9477†	0.9556†	**0.9251**†
oracle	0.9587	0.9559	0.9669	0.9654	0.9675	0.9760	0.9085

Table 4.5: The accuracy for the target domains of type I test. The boldface indicates the best performance and the underline marks the second best result. $\sim$: p-value $\geq$ 0.05, $\dagger$: p-value $\ll$ 0.05 with regard to the baseline output. The background is color-coded in the same way as Table 4.1.

competing models. Similar to the previous outcome, the disease status like diabetes also has a big impact on the vessel segmentation performance, likely due to the presence of retinal hemorrhages and hard exudates. The accuracy, recall and Hausdorff distance results show similar patterns.

4.3.5 Discussion

In this work, I propose a DG approach that includes a novel data augmentation method, a shape description via Hessian-based vector field and a transformer architecture that takes advantage of the multi-scale local feature. Experimentally, I show that all these three components contribute to the superior DG performance of my approach.

Unlike traditional data augmentations that modify the image in terms of quality, appearance and spatial configuration, such as suggested in (Zhang et al., 2020b), I use the anatomy-consistent pseudo-modalities generated with the synthesis model proposed in Section 3.5. The advantage of this method is that the vessel structure is fixed while the style of the image remains flexible, allowing me to use it for random data augmentation. I use the tubular shape of the vessel, which provides a domain-agnostic feature, to align the data distribution. Inspired by the Frangi filter (Frangi et al., 1998), the delineation of the vessel shape is achieved by the eigenvector of the Hessian matrix. Ideally, the vector field will mimic the blood flow along the vessels.

Method	ARIA		HRF			OCTA-500	ROSE
	amd	diabetic	healthy	diabetic	glaucoma		
baseline	0.6305	0.6580	0.5454	0.5064	0.5202	**0.8745**	0.6244
+Regularizer	0.6190	0.6499	0.5417	0.4904	0.5189	0.8088	**0.6542**
+BigAug	**0.7317**	**0.7692**	0.6148	0.5773	0.6049	0.8614	0.6363
+MASF	0.6078	0.6558	0.4567	0.4350	0.4509	0.7557	0.5612
Proposed	0.6390$^{\sim}$	0.6599$^{\sim}$	**0.6361**†	**0.5841**†	**0.6523**†	0.8620†	0.6535
oracle	0.7435	0.6948	0.8716	0.7803	0.8089	0.8466	0.9551

Table 4.6: The recall for the target domains of type I test. The boldface indicates the best performance and the underline marks the second best result. $^{\sim}$: p-value $\geq$ 0.05, † : p-value $\ll$ 0.05 with regard to the baseline output. The background is color-coded in the same way as Table 4.1.

Method	ARIA		HRF			OCTA-500	ROSE
	amd	diabetic	healthy	diabetic	glaucoma		
baseline	0.0699	0.0761	0.0369	0.0436	**0.0352**	0.0590	0.0520
+Regularizer	0.0746	0.0771	0.0372	0.0436	0.0366	0.0556	0.0512
+BigAug	0.0726	0.0806	**0.0361**	**0.0426**	0.0353	0.0584	0.0670
+MASF	**0.0688**	**0.0724**	0.0403	0.0473	0.0456	0.0533	**0.0492**
Proposed	0.0780†	0.0839†	0.0366†	0.0429†	0.0362†	**0.0506**†	0.0560†
oracle	0.0631	0.0696	0.0335	0.0386	0.0313	0.0503	0.0497

Table 4.7: The normalized Hausdorff distance for the target domains of type I test. The boldface indicates the best performance and the underline marks the second best result. $^{\sim}$: p-value $\geq$ 0.05, † : p-value $\ll$ 0.05 with regard to the baseline output. The background is color-coded in the same way as Table 4.1.

I attempt to connect the traditional model-based algorithm with the deep learning so that the model can be developed based on the prior of human knowledge instead of a purely data-driven approach. In addition to the improvement of generalization, this is a method that can potentially make the black-box of deep models more interpretable for users. Lastly, I design a model architecture that is suitable for the vector field input. Unlike natural images with large scale features, medical images usually stress more on local features. Therefore, the proposed VFT computes the attention within the parcellation with three different patch sizes.

From the experiments, I observe that although VFT is relatively robust on domain shift, its performance can be greatly impacted by the contrast condition of the image as it heavily relies on the gradients. On datasets with very poor contrast, the segmentation result is not improved. Another limitation is that VFT has longer inference time since it requires to compute the Hessian and eigenvectors which can be time consuming. Hence, VFT is not a good option for real-time vessel segmentation. However, the vector field has advantage in interpretation as it is aligned with human perception of vessel. The future work can be using reinforcement learning to trace the vessels given the vector field.

In conclusion, my method significantly enhances the generalizability for vessel segmentation in cross-resolution and cross-modality situations. Further improvements are needed for optimizing the performance across disease status. Although the current work is conducted on 2D vessel segmentation, the idea of creating

Method	DRIVE	STARE		ARIA			HRF	
			health	amd	diabetic	health	diabetic	glaucoma
baseline	0.4635	0.4454	0.5207	0.4529	0.4429	0.5774	0.4489	0.5230
+Regularizer	0.5562	0.5436	**0.5650**	**0.5075**	**0.5224**	0.5747	0.4533	0.4985
+BigAug	0.5483	0.5121	0.5650	0.5036	0.5210	0.6078	0.4699	0.5423
+MASF	0.4855	0.4524	0.5309	0.4633	0.4529	0.5871	0.4587	0.5326
Proposed	**0.5753**†	**0.5576**†	0.5395$^\sim$	0.4660$^\sim$	0.4812$^\sim$	**0.6263**†	**0.5187**†	**0.5736**†
oracle	0.8227	0.8154	0.7448	0.7334	0.7065	0.8358	0.7524	0.7732

Table 4.8: The Dice values for the target domains of type II test. The boldface indicates the best performance and the underline marks the second best result. $^\sim$: p-value $\geq$ 0.05, † : p-value $\ll$ 0.05 with regard to the baseline output. The background is color-coded in the same way as Table 4.1.

Method	DRIVE	STARE		ARIA			HRF	
			health	amd	diabetic	health	diabetic	glaucoma
baseline	0.9418	0.8991	0.9301	**0.9260**	0.9223	0.9381	0.9362	0.9445
+Regularizer	0.9508	0.9123	0.9230	0.9218	0.9263	0.9315	0.9270	0.9330
+BigAug	0.9503	0.9105	0.9271	0.9221	0.9271	0.9330	0.9230	0.9333
+MASF	0.9522	0.9135	**0.9304**	0.9235	**0.9350**	0.9411	0.9304	0.9335
Proposed	**0.9587**†	**0.9160**†	0.9197†	0.9119†	0.9132†	**0.9427**†	**0.9381**$^\sim$	**0.9468**$^\sim$
oracle	0.9683	0.9629	0.9577	0.9587	0.9559	0.9669	0.9654	0.9675

Table 4.9: The accuracy values for the target domains of type II test. The boldface indicates the best performance and the underline marks the second best result. $^\sim$: p-value $\geq$ 0.05, † : p-value $\ll$ 0.05 with regard to the baseline output. The background is color-coded in the same way as Table 4.1.

the shape-aware representation can be readily propagated to 3D or other applications in future research. An early version was presented at *International Conference on Medical Imaging with Deep Learning (2022)*. This work is published on *Medical Image Analysis (2024)*. The code and deep model checkpoints are available at https://github.com/MedICL-VU/Vector-Field-Transformer.

Method	DRIVE	STARE		ARIA			HRF	
			health	amd	diabetic	health	diabetic	glaucoma
baseline	0.4216	0.3752	0.4600	0.4095	0.3997	0.4637	0.4023	0.4839
+Regularizer	0.5743	0.5029	0.5759	0.5426	0.5513	0.5181	0.4653	0.5068
+BigAug	0.5386	**0.5262**	**0.5829**	**0.5476**	**0.5607**	**0.5803**	0.5183	**0.5980**
+MASF	0.5203	0.4899	0.5604	0.5332	0.5065	0.5402	0.4967	0.5341
Proposed	**0.5813**†	0.5187†	0.5584†	0.5121†	0.5159†	0.5413†	**0.5187**†	0.5589†
oracle	0.8384	0.7793	0.7387	0.7435	0.6948	0.8716	0.7803	0.8089

Table 4.10: The recall for the target domains of type II test. The boldface indicates the best performance and the underline marks the second best result. $^\sim$: p-value ≥ 0.05, † : p-value $\ll 0.05$ with regard to the baseline output. The background is color-coded in the same way as Table 4.1.

Method	DRIVE	STARE		ARIA			HRF	
			health	amd	diabetic	health	diabetic	glaucoma
baseline	0.0603	0.0639	0.0734	0.0800	0.0888	**0.0369**	**0.0436**	**0.0361**
+Regularizer	0.0558	0.0607	**0.0704**	**0.0831**	**0.0831**	0.0379	0.0445	0.0370
+BigAug	0.0538	0.0652	0.0733	0.0792	0.0860	0.0376	0.0450	0.0373
+MASF	0.0550	**0.0574**	0.0735	0.0854	0.0881	0.0377	0.0448	0.0374
Proposed	0.0531†	0.0615†	0.0732†	0.0818†	0.0862†	0.0372†	0.0444†	0.0367†
oracle	0.0378	0.0494	0.0571	0.0631	0.0696	0.0335	0.0386	0.0313

Table 4.11: The normalized Hausdorff distance for the target domains of type II test. The boldface indicates the best performance and the underline marks the second best result. $^\sim$: p-value ≥ 0.05, † : p-value $\ll 0.05$ with regard to the baseline output. The background is color-coded in the same way as Table 4.1.

4.4 Explicit shape modelling: VesselMorph

Based on the vector field transformer, I propose a novel method, *VesselMorph*, to merge the Hessian-based shape description (Hu et al., 2021b) with the principles of diffusion tensor imaging (DTI) (Le Bihan et al., 2001). I introduce a bipolar tensor field (BTF) to explicitly represent the vessel shape by a tensor at each pixel. To effectively merge the features in the intensity image and the shape descriptor BTF, I employ a full-resolution feature extraction network to obtain an interpretable representation in the latent space from both inputs. This technique is broadly used in unsupervised segmentation (Liu et al., 2020; Hu et al., 2021a) and representation disentanglement (Dewey et al., 2020; Ouyang et al., 2021). I also leverage this type of model to generate the pseudo-modalities in Section 3.5.

As shown in Figure 4.9, let $\mathbf{x}$ be the input image and $\Psi(\mathbf{x})$ the corresponding BTF. $D(E^I(\cdot))$ and $D(E^S(\cdot))$ are two feature extraction networks with a shared decoder D. I empirically observe that the intensity representation $\mathbf{z}^I$ can precisely delineate thinner vessels while the structure representation $\mathbf{z}^S$ works better on thick ones. I combine the strengths of the two pathways for a robust DG performance. The two latent images are fused by a weight-balancing trick $\Gamma(\mathbf{z}^I, \mathbf{z}^S)$ to avoid any potential bias induced by the selection of source domains. Finally, I train a segmentation network D^T on the fused latent images. I compare the performance of VesselMorph to other DG approaches on six public datasets that represent various distribution shift conditions, and show that VesselMorph has superior performance in most OOD domains.

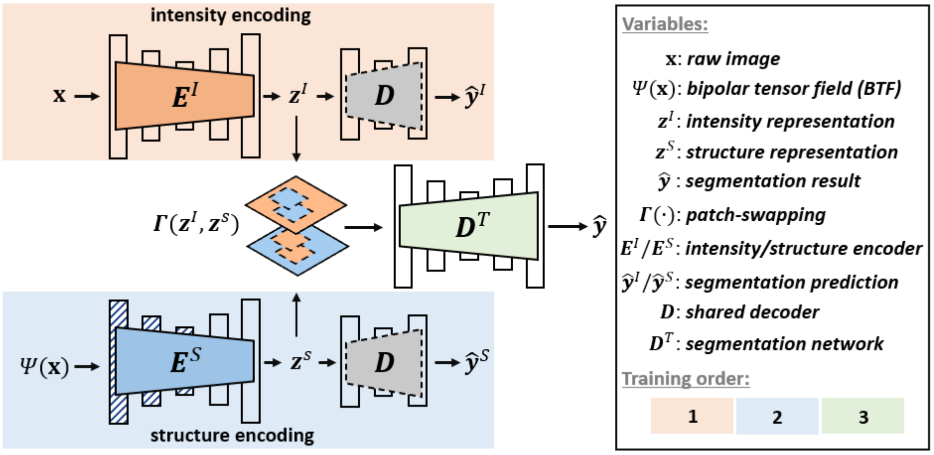

intensity encoding
x
E^I
z^I
D
$\hat{y}^I$
$\Gamma(z^I, z^S)$
D^T
$\hat{y}$
$\Psi(\mathbf{x})$
E^S
z^s
D
$\hat{y}^S$
structure encoding
Variables:
$\mathbf{x}$: raw image
$\Psi(\mathbf{x})$: bipolar tensor field (BTF)
z^I: intensity representation
z^S: structure representation
$\hat{y}$: segmentation result
$\Gamma(\cdot)$: patch-swapping
E^I/E^S: intensity/structure encoder
$\hat{y}^I/\hat{y}^S$: segmentation prediction
D: shared decoder
D^T: segmentation network
Training order:
1
2
3

Figure 4.9: The overall model structure of VesselMorph. The shaded layers include transformer blocks. The dashed line indicates D will be discarded in testing.

4.4.1 Method

4.4.1.1 Bipolar Tensor Field

Unlike ML models, human's visual interpretation of vessels is rarely affected by data distribution shifts. Mimicking the human vessel recognition can thus help address the DG problem. In addition to intensity

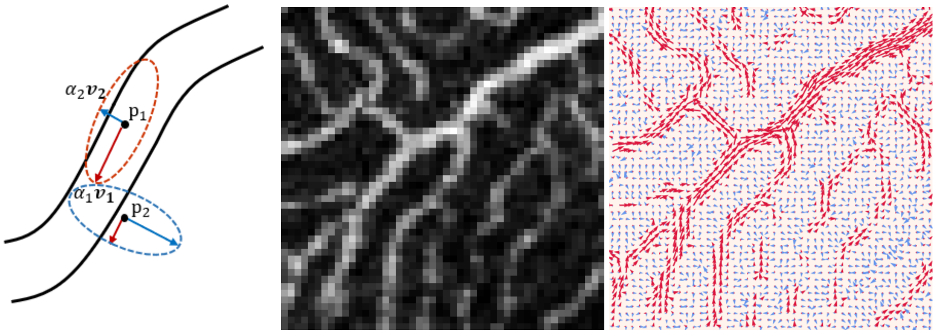


Figure 4.10: **Left:** A simplified illustration of BTF. The red arrows indicate the orientation of $\mathbf{v}_1$ while the blue arrows correspond to $\mathbf{v}_2$. The ellipses represent the tensors at p_1 (in the vessel) and p_2 (in the background). **Right:** BTF applied on an OCTA image.

values, human perception of vessels also depends on the local contrast and the correlation in a neighborhood, which is often well described by the local Hessian. Inspired by the use of DTI to depict the white matter tracts, I create a Hessian-based bipolar tensor field to represent the morphology of vessels. Here, I leverage the same method as described in Equation 4.3 to compute the 2×2 Hessian matrix $\mathscr{H}(\mathbf{x}(i,j),\sigma^*)$ at pixel (i,j). The optimal standard deviation σ^* is acquired with grid search in a range $[\sigma_{min},\sigma_{max}]$ to maximize the vesselness $\mathscr{V}(\sigma)$ defined in Equation 4.2. Still, I apply the eigen decomposition to obtain the eigenvalues λ_1, λ_2 ($|\lambda_1| \leq |\lambda_2|$) and the corresponding eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$ at the optimal σ^*.

Instead of solely analyzing the signs and magnitudes of the Hessian eigenvalues as in the traditional Frangi filter (Frangi et al., 1998), I propose to leverage the eigenvectors along with custom-designed magnitudes to create the tensor field as shown in Figure 4.10 (Left). The core idea of Frangi filter is to enhance the tubular structure by matching the vessel diameter with the distance between the two zero crossings in the second order derivative of Gaussian ($2\sqrt{2}\sigma^*$). However, the solution is not guaranteed to land in range $[\sigma_{min},\sigma_{max}]$, especially for small vessels. Consequently, I observe that the inaccurate estimation of σ^* results in a blurring effect at the vessel boundary, which is problematic for segmentation. As an example in Fig. 4.10 (Left), the direction of $\mathbf{v}_1$ at p_2 aligns with that at p_1, even though p_1 is inside the vessel while p_2 is in the background but close to the boundary. This makes it difficult for the vector orientations alone to differentiate points inside and outside the vessel. To tackle this, I introduce the idea of a bipolar tensor by assigning a large magnitude to the orthogonal eigenvector $\mathbf{v}_2$ to points in the background, as shown in the blue dashed ellipse. Specifically, I define the magnitudes α_1 and α_2 associated with the eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ as:

$$\alpha_1 = \underbrace{P(\mathbf{x} \leq \mathbf{x}_{ij})}_{\text{bright}}\underbrace{\exp\left(-\varepsilon\frac{\lambda_1^2}{\|\mathscr{H}\|_F^2}\right)}_{\text{vessel-like}}, \quad \alpha_2 = \underbrace{P(\mathbf{x} > \mathbf{x}_{ij})}_{\text{dark}}\underbrace{\exp\left(-\varepsilon\frac{\lambda_2^2}{\|\mathscr{H}\|_F^2}\right)}_{\text{vessel-like}} \tag{4.4}$$

where $P(\mathbf{x} > \mathbf{x}_{ij})$ is the probability that the intensity of a random pixel x in the image is greater than $\mathbf{x}_{ij}$. This is equivalent to normalizing the histogram by the factor hw and computing the cumulative distribution function at $\mathbf{x}_{ij}$. This term thus provides a normalized brightness function in the range $[0,1]$. The exponential term represents how vessel-like the pixel is by using a normalized eigenvalue, and is in the $[0,1]$ range as well. ε is a constant that controls the sensitivity, which is empirically set to 0.5. With the custom magnitudes α_1 and α_2, the two poles can better differentiate vessels from the background. Figure 4.10 (Right) is an example of BTF on an OCTA image. In practice, I stack the two vectors as the input to the structural encoding network, i.e., $\Psi(\mathbf{x}_{ij}) = \left[\alpha_1\mathbf{v}_1^\top, \alpha_2\mathbf{v}_2^\top\right]^\top \in \mathbb{R}^{4\times 1}$.

4.4.1.2 Latent Vessel Representation

Preserving the spatial resolution for the bottleneck of models with U-Net backbone is a common strategy to emphasize the structural features in unsupervised segmentation (Liu et al., 2020; Hu et al., 2021a) and representation disentanglement (Dewey et al., 2020; Ouyang et al., 2021). I employ a network that has a full-resolution ($H \times W$ pixels) latent space as the feature extraction model. I propose to extract vessel structure from both the intensity image $\mathbf{x} \in \mathbb{R}^{H\times W}$ and its corresponding BTF, $\Psi(\mathbf{x}) \in \mathbb{R}^{4\times H\times W}$. Therefore, in Figure 4.9, the intensity $D(E^I(\cdot))$ and structure $D(E^S(\cdot))$ encoding pathways share the decoder D, and the latent images $\mathbf{z}^I, \mathbf{z}^S \in \mathbb{R}^{H\times W}$. To distribute more workload on the encoder, D has a shallower architecture and will be discarded in testing. For the intensity encoding, the model is optimized by minimizing the segmentation loss function defined as the combination of cross-entropy and Dice loss:

$$\mathscr{L}_{seg} = \underbrace{-\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{y}(i,j)\log\tilde{\mathbf{y}}(i,j)}_{\text{cross entropy loss}} + \underbrace{\left(1 - \frac{2\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{y}(i,j)\tilde{\mathbf{y}}(i,j)}{\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{y}(i,j)^2 + \tilde{\mathbf{y}}(i,j)^2}\right)}_{\text{Dice loss}} \tag{4.5}$$

where H and W are the height and width of the image, $\mathbf{y}$ is the ground truth and $\hat{\mathbf{y}}^I$ is the prediction from the training-only decoder D. Although there is no explicit constraint on the latent image $E^I(\mathbf{x}) = \mathbf{z}^I$, I note that the segmentation-based supervision encourages it to include the vessels while most other irrelevant features are filtered out. Hence, I can view the latent feature as a vessel representation.

My approach is slightly different for the structure encoding as I notice that it is hard for the feature extraction network to generate a stable latent image that is free of artifacts when the number of input channels is greater than 1. Thus, it is necessary to use E^I as a teacher model that provides direct supervision on the vessel representation. In other words, I first train the intensity encoding path to get E^I and D, then train the E^S by leveraging both the segmentation loss in Equation 4.5 and a similarity loss defined as:

$$\mathscr{L}_{sim}(\mathbf{z}^S, \mathbf{z}^I) = \sum_{n=1}^{N} \|\mathbf{z}_n^S - \mathbf{z}_n^I\|_1 + \mathrm{SSIM}(\mathbf{z}^S, \mathbf{z}^I) \tag{4.6}$$

where $N = H \times W$ is the total number of pixels in the image. The structural similarity loss SSIM (Hore and Ziou, 2010) between image A and image B is defined as:

$$\mathrm{SSIM}(A,B) = 2\frac{(2\mu_A\mu_B + c_1)(2\sigma_{AB} + c_2)}{(\mu_A^2 + \mu_B^2 + c_1)(\sigma_A^2 + \sigma_B^2 + c_2)} \tag{4.7}$$

where μ and σ represent the mean and standard deviation of the image, and I set $c_1 = 0.01$ and $c_2 = 0.03$. The overall loss function for the structural encoding is thus a weighted sum of $\mathscr{L}_{seg}$ and $\mathscr{L}_{sim}$:

$$\mathscr{L}(\Psi(\mathbf{x}), \mathbf{y}) = \omega_1 \mathscr{L}_{seg}(\hat{\mathbf{y}}^S, \mathbf{y}) + \omega_2 \mathscr{L}_{sim}(\mathbf{z}^S, \mathbf{z}^I) \tag{4.8}$$

Empirically, I set the weights $\omega_1 = 1$, $\omega_2 = 5$. Experimentally, I found that the $\mathbf{z}^I$ is good at preserving small vessels, while $\mathbf{z}^S$ works better on larger ones.

4.4.1.3 Fusion of Vessel Representations

Given the two synthesized vessel representations $\mathbf{z}^I$ and $\mathbf{z}^S$, I need to introduce a fusion method to take advantage of both intensity and structure features. Naively stacking these two channels as input to the segmentation network is prone to inducing bias: if $\mathbf{z}^I$ is consistently better for images from the source domain, then the downstream task model D^T would learn to downplay the contribution of $\mathbf{z}^S$ due to this biased training data. As a result, despite its potential to improve performance, $\mathbf{z}^S$ would be hindered from making a significant contribution to the target domain during testing. To circumvent this issue, I propose a simple weight-balancing trick. As illustrated in Figure 4.9, I randomly swap some patches between the two latent images so that D^T does not exclusively consider the feature from a single channel, even for biased training data. This trick is feasible because $\mathbf{z}^S$ and $\mathbf{z}^I$ are in the same intensity range, due to the similarity constraints applied in Equation 4.6. Thus the input to D^T is $\tilde{\mathbf{x}} = \Gamma(\mathbf{z}^I, \mathbf{z}^S)$, where $\tilde{\mathbf{x}} \in \mathbb{R}^{2\times H\times W}$. The loss function leveraged for D^T is the same as Equation 4.5.

The complete algorithm for training of VesselMorph is shown in Algorithm 1. Briefly, I first train the intensity encoder E^I as it is easier to generate a stable vessel representation $\mathbf{z}^I$. Then a structure encoder E^S is trained with the supervision of the ground truth and teacher model E^I so that an auxiliary representation $\mathbf{z}^S$

is extracted from the structural descriptor BTF. The last step is to train a segmentation network D^T with the fusion of the two vessel maps $\Gamma(\mathbf{z}^I, \mathbf{z}^S)$. During testing, the patch-swapping is no longer needed, so I simply concatenate $E^I(\mathbf{x})$ and $E^S(\Psi(\mathbf{x}))$ as the input to D^T.

Algorithm 1: Training of VesselMorph

```
input  : Source domains $\mathscr{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{K}$,
         hyperparameters: $\varepsilon, \sigma_{min}, \sigma_{max}$, $\eta_{E^I}$, $\eta_{E^S}$, $\eta_{D^T}$ $\omega_1$, $\omega_2$
output: parameters of models $\theta_I^*$, $\theta_S^*$, $\varphi_T^*$
// Train the intensity encoder $E^I$ as a teacher model
repeat
    for $i = 1 : K$ do
        $\theta_I^{'} \leftarrow \theta_I - \eta_{E^I}(i)\nabla\mathscr{L}_{seg}(D(E^I(\mathbf{x}_i)), \mathbf{y}_i)$
until converge
Generate the tensor field: $\Psi(\mathbf{x})$
// Train the structure encoder $E^S$ as a student model
repeat
    for $i = 1 : K$ do
        $\hat{\mathbf{y}}_i \leftarrow D(E^S(\Psi(\mathbf{x}_i)))$
        $\mathscr{L}(\Psi(\mathbf{x}_i), \mathbf{y}_i) \leftarrow \omega_1\mathscr{L}_{seg}(\hat{\mathbf{y}}_i, \mathbf{y}_i) + \omega_2\mathscr{L}_{sim}(E^S(\Psi(\mathbf{x}_i)), E^I(\mathbf{x}_i))$
        $\theta_S^{'} \leftarrow \theta_S - \eta_{E^S}(i)\nabla\mathscr{L}(\Psi(\mathbf{x}_i), \mathbf{y}_i)$
until converge
// Train the segmentation network $D^T$
repeat
    for $i = 1 : K$ do
        $\tilde{\mathbf{x}}_i \leftarrow \Gamma(E^I(\mathbf{x}_i), E^S(\Phi(\mathbf{x}_i)))$
        $\varphi_T^{'} \leftarrow \varphi_T - \eta_{D^T}(i)\nabla\mathscr{L}_{seg}(D^T(\tilde{\mathbf{x}}_i), \mathbf{y}_i)$
until converge
```

4.4.2 Datasets

The 6 publicly available datasets used in this study are listed in Table 4.12. Since there are more labeled fundus data available, I set up a source domain $\mathscr{S}$ that includes three fundus datasets: DRIVE, STARE and the control subjects in ARIA. In the target domain $\mathscr{T}$, I test the performance of the model under three different conditions: pathology (diabetic/AMD subjects in ARIA), resolution change (HRF) and cross-modality (OCTA500 and ROSE).

	modality	**resolution**	**# sample**	**domain**
DRIVE Staal et al. (2004)	fundus	565×584	20	$\mathscr{S}$
STARE Hoover et al. (2000)	fundus	700×605	20	$\mathscr{S}$
ARIA Farnell et al. (2008)	fundus	768×576	61/59/23	$\mathscr{S}/\mathscr{T}/\mathscr{T}$
HRF Budai et al. (2013)	fundus	3504×2336	15/15/15	$\mathscr{T}$
OCTA-500(6M) Li et al. (2020b)	OCTA	400×400	300	$\mathscr{T}$
ROSE Ma et al. (2020)	OCTA	304×304	30	$\mathscr{T}$

Table 4.12: Publicly available datasets used in my experiments. For ARIA and HRF, I list the number of samples per class. ARIA classes: healthy, diabetic and AMD (age-related macular degeneration). HRF classes: healthy, diabetic and glaucoma. The shading of the rows indicates datasets in similar distributions to each other.

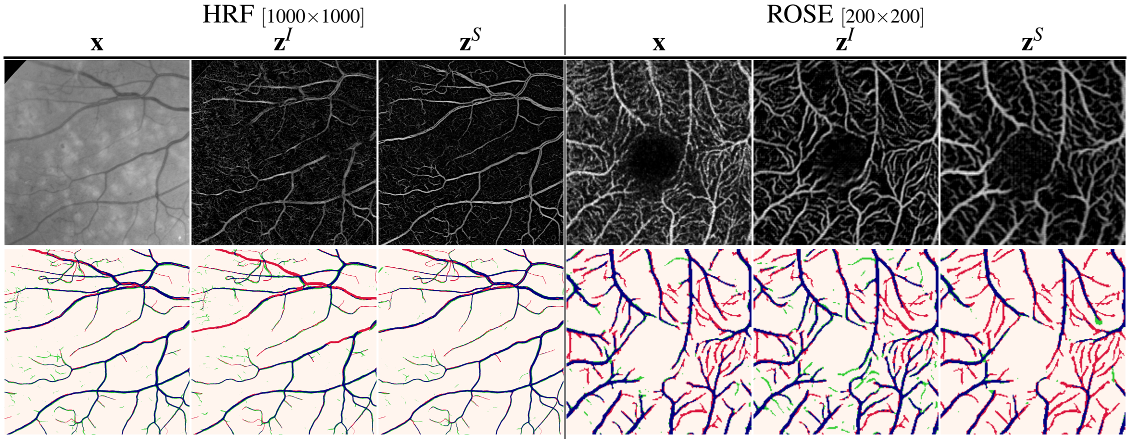


Figure 4.11: Qualitative ablation. The shown patches are 1000×1000pix for HRF diabetic image and 200×200pix for ROSE. **Top row:** raw image, $\mathbf{z}^I$ and $\mathbf{z}^S$. **Bottom row:** the VesselMorph segmentation and prediction from each pathway, i.e., $D^T(\Gamma(\mathbf{z}^I, \mathbf{z}^S))$, $D(\mathbf{z}^I)$, and $D(\mathbf{z}^S)$. **Red** and **green** indicate the false negative (FN) and false positive (FP), respectively. $\mathbf{z}^I$ may miss large vessels, while $\mathbf{z}^S$ may miss thin ones.

4.4.3 Experiments

4.4.3.1 Compared Methods

Similar with what I did in Section 4.3, I select one representative algorithm from each of the three major categories of DG approaches as a competing method. For data augmentation, I implement BigAug (Zhang et al., 2020b). For meta-learning, I use the MASF (Dou et al., 2019) model. For domain alignment, I use the domain regularization network (Aslani et al., 2020). In addition, I also include my model VFT (Hu et al., 2021b) presented in the previous section. The baseline model is a vanilla residual U-Net trained on $\mathscr{S}$, and the oracle model is the same network trained directly on each target domain to represent the optimal performance. Note that for a fair comparison, I set the baseline model to have a bit more parameters than $D(E^I(\cdot))$ $(7.4 \times 10^5 : 6.7 \times 10^5)$.

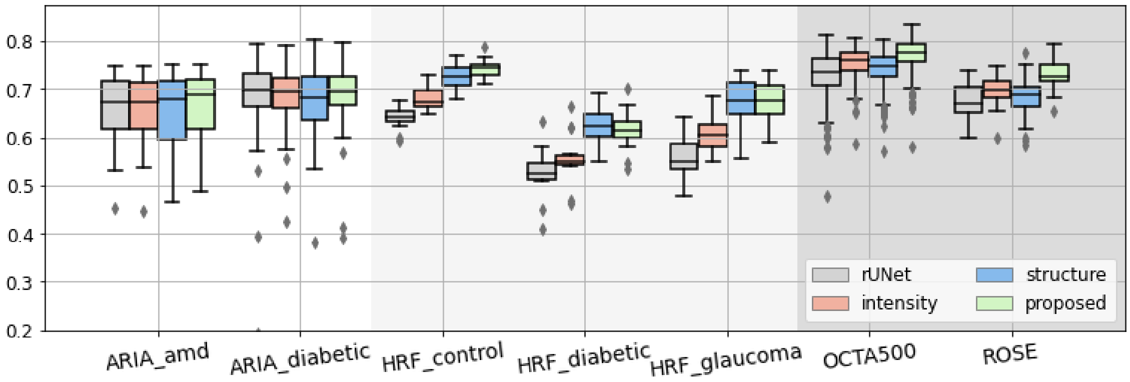


Figure 4.12: Quantitative ablation results. Dice scores of vanilla residual U-Net, intensity encoding $D(\mathbf{z}^I)$, structural encoding $D(\mathbf{z}^S)$ and the final output $D^T(\Gamma(\mathbf{z}^I, \mathbf{z}^S))$.The background is encoded the same way as Table 1. I note that the $\mathbf{z}^S$ is especially useful in capturing the thick vessels in HRF, whereas $\mathbf{z}^I$ provides additional precision in the thin vessels in the OCTA datasets. The proposed model combines these advantages and is robust across the board.

4.4.3.2 Implementation Details

I leverage the residual U-Net structure for E^I, D and D^T. To take advantage of the tensor field, the structure encoder E^S is equipped with parallel transformer blocks with different window sizes as proposed in VFT (Hu et al., 2021b). All networks are trained and tested on an NVIDIA RTX 2080TI 11GB GPU. I use a batch size of 5 and train for 100 epochs. I apply the Adam optimizer with the initial learning rate $\eta_{E^I} = \eta_{E^S} = 5 \times 10^{-4}$, $\eta_{D^T} = 1 \times 10^{-3}$, decayed by 0.5 for every 3 epochs. For fundus images, I set the green channel as network input $\mathbf{x}$. The intensity values are normalized to $[0, 1]$.

4.4.4 Results

Figure 4.11 shows a qualitative ablation study: it illustrates that the intensity representation $\mathbf{z}^I$ may miss large vessels in the very high-resolution HRF images, while $\mathbf{z}^S$ remains robust. In contrast, $\mathbf{z}^I$ provides sharper delineation for very thin vessels in ROSE. The fusion of both pathways outperforms either pathway for most scenarios. These observations are further supported by the quantitative ablation study in Figure 4.12. I note that $\mathbf{z}^S$ and $\mathbf{z}^I$ can be used as synthetic angiograms that provide both enhanced vessel visualization and model interpretability.

Figure 4.13 shows the t-SNE plots (Van der Maaten and Hinton, 2008) of the datasets. The distribution gaps between datasets are greatly reduced for the two latent vessel representations.

Table 4.13 compares all methods on the target domain $\mathscr{T}$. For the diseased ARIA data, all methods show comparable performance and are not significantly different from the baseline. VesselMorph has the best OOD outcome for both cross-modality (dark gray) and cross-resolution (light gray) scenarios, except the OCTA500 dataset where VFT, MASF and VesselMorph perform similarly. The results of VFT and VesselMorph prove

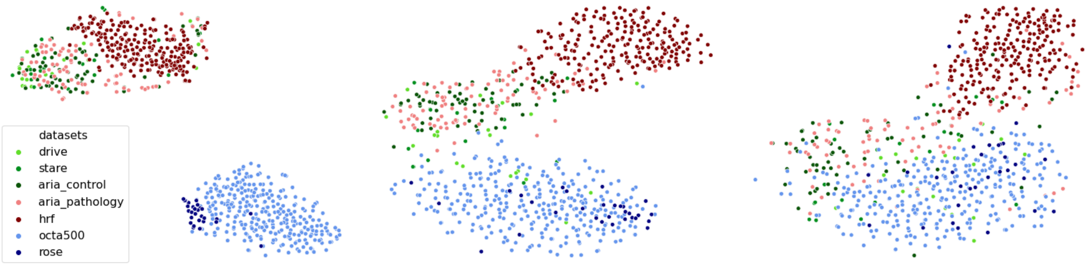


Figure 4.13: t-SNE on raw data $\mathbf{x}$(left), $\mathbf{z}^I$(center) and $\mathbf{z}^S$(right). $\mathscr{S}$ is coded by shades of green, while fundus and OCT-A in $\mathscr{T}$ are coded by red and blue shades respectively. Both intensity and structure representations reduce the domain gaps between datasets.

Method	ARIA		HRF			OCTA 500	ROSE
	amd	diabetic	healthy	diabetic	glaucoma		
baseline	0.6382	0.6519	0.6406	0.5267	0.5566	0.7316	0.6741
+Regular	0.6489	0.6697	0.6403	0.5216	0.5625	0.7354	0.6836
+BigAug	0.6555	0.6727	0.6613	0.5389	0.5735	0.7688	0.6932
+MASF	0.6533	0.6775	0.6131	0.5358	0.5629	0.7765	0.6725
VFT	0.6181	0.6405	0.7058	0.5732	0.6410	**0.7791**	0.7281
VesselMorph	**0.6619**$^{\sim}$	**0.6787**$^{\sim}$	**0.7420**†	**0.6145**†	**0.6756**†	0.7714†	**0.7308**†
oracle	0.7334	0.7065	0.8358	0.7524	0.7732	0.8657	0.7603

Table 4.13: Dice values for testing on target domains. **Boldface**: best result. Underline: second best result. $^{\sim}$: p-value ≥ 0.05, † : p-value $\ll 0.05$ in paired t-test against the baseline output. The background is color-coded the same way as Table 4.12.

the value of the shape information.

4.4.5 Discussion

In this work, I propose to solve the DG problem by explicitly modeling the domain-agnostic tubular vessel shape with a bipolar tensor field which connects traditional algorithms with deep learning. I extract vessel representation from both intensity and BTF, then fuse the information from the two pathways so that the segmentation network can better exploit both types of description. My VesselMorph model provides significant quantitative improvement on Dice score across a variety of domain shift conditions, and its latent images offer enhanced vessel visualization and interpretability.

4.5 Implicit shape modelling: MAP

In Section 4.3 and Section 4.4, I propose to explicitly delineate the vessel shape by a Hessian-based vector/tensor field. However, the dependency on the image gradient makes this approach vulnerable to low-quality data with poor contrast and/or high noise. In this section, I instead propose an implicit way of exploiting the morphological features by adopting the **m**eta-learning paradigm on **a**natomy-consistent **p**seudo-modalities (MAP).

First, I leverage a structural feature extraction network to generate three different anatomy-consistent pseudo-modalities. This method has been discussed in detail in Section 3.5. The synthesis model is again shown in Figure 4.14(a) and the three pseudo-modalities are denoted as $\mathscr{D}^1$, $\mathscr{D}^2$ and $\mathscr{D}^3$ (Figure 4.14(b)).

Meta-learning has recently emerged as a popular technique for addressing the DG problem (Dou et al., 2019; Khandelwal and Yushkevich, 2020). Following the idea of episodic training presented in MAML (Finn et al., 2017), researchers split their training data into two subsets, meta-train and meta-test, to mimic the scenario of encountering out-of-distribution (OOD) data during training. Liu et al. (2021) proposed to conduct meta-learning in a continuous frequency space created by mixing up (Zhang et al., 2017; Kim et al., 2020) the amplitude spectrum. They keep the phase spectrum unchanged to preserve the anatomy in the generated images. In contrast, given the pseudo-modalities with identical underlying vasculature, I am able to create a continuous image space via Dirichlet mixup (Shu et al., 2021) without affecting the vasculature. I regard images in each pseudo-modality as a corner of a tetrahedron, as depicted in Figure 4.14(c). The red facet of the tetrahedron is a continuous space created by the convex combination of images from the three pseudo-modalities. I use images in one pseudo-modality (blue node) for meta-train and the mixup space (red facet) for meta-test. An important property of the mixup space is that all the samples share the same vessel structure while the image style may differ drastically. Hence, employing proper constraints on the relationship between features can implicitly encourage the model to learn the shape of vessels. Inspired by the idea presented in (Dou et al., 2019), I leverage a similarity loss to express the feature consistency between the meta-train and meta-test stages. Additionally, I propose a normalized cross-correlation (NCC) loss to differentiate latent features extracted from images with different anatomy. In the context of contrastive learning, these loss functions cluster positive pairs and separate negative pairs.

In Section 4.1, three types of domain shift are described, type I: pathological phenotypes, type II: cross-site shift and type III: cross-modality shifts (note that these are different than the two types of Source/Target domain settings in Section 4.3). The proposed method will be tested on all these three scenarios to comprehensively evaluate its DG performance. Specifically, I use seven public datasets including color fundus, OCT angiography (OCT-A) and fluorescein angiography (FA) images. The MAP is trained on fundus data and

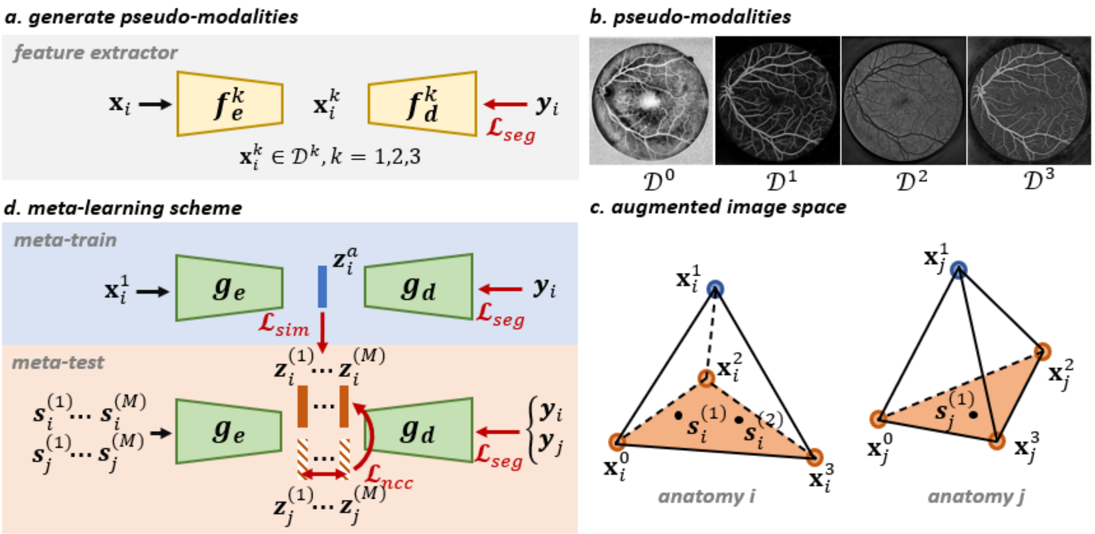


Figure 4.14: The key components of MAP, clockwise. **(a)** $f(\cdot)$ is the synthesis network. $\mathbf{x}_i$ is the i^{th} color fundus input and $\mathbf{y}_i$ is its ground truth vessel map. k indexes three different models that generate diverse pseudo-modalities. **(b)** An example image in four pseudo-modalities: $\mathscr{D}^0$ is the histogram equalization of intensity-reversed green channel of input $\mathbf{x}$ and $\mathscr{D}^k$, $k = 1,2,3$ are generated by f_e^k. **(c)** The four pseudo-modalities of an input $\mathbf{x}_i$ form the corners of a tetrahedron. The colored facet is a continuous image space created by Dirichlet mixup. $\mathbf{s}_i^{(m)}$ denotes the m^{th} sample from the image space. Anatomy i represents the underlying shape of vasculature in $\mathbf{x}_i$, which is consistent for all samples $\mathbf{s}_i^{(m)}$. **(d)** The meta-learning scheme. $g(\cdot)$ is the segmentation network, M is the number of samples drawn, $\mathbf{z}$ is the latent feature vector.

tested on all modalities. I show that MAP exhibits outstanding generalization ability in most conditions.

4.5.1 Method

4.5.1.1 Pseudo-modality Synthesis

Since the method used for pseudo-modality synthesis has been thoroughly presented in Section 3.5, here I just clarify the notations for models and variables in this study. I denote the input image and the corresponding ground truth vessel map with $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ where the subscript i is the index, and N is the total number for the paired data. In Figure 4.14(a), the superscript $k = 1,2,3$ indexes the three different pseudo-modalities. These modalities are generated by distinct models, i.e., $f_e^k(\mathbf{x}_i) = \mathbf{x}_i^k$. I define each of them as an individually distributed domain $\mathscr{D}^1$, $\mathscr{D}^2$ and $\mathscr{D}^3$. In other words, $\mathscr{D}^k = \{\mathbf{x}_i^k, \mathbf{y}_i\}_{i=1}^N$. Note that I assume the label $\mathbf{y}$ is consistent across diverse data distribution. The segmentation loss $\mathscr{L}_{seg}$ in both Figure 4.14(a) and (d) is set to be the sum of cross-entropy and Dice loss (Ma et al., 2021), i.e.,

$$\mathscr{L}_{seg} = \underbrace{-\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{y}(i,j)\log\tilde{\mathbf{y}}(i,j)}_{\text{cross entropy loss}} + \underbrace{\left(1-\frac{2\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{y}(i,j)\tilde{\mathbf{y}}(i,j)}{\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{y}(i,j)^2+\tilde{\mathbf{y}}(i,j)^2}\right)}_{\text{Dice loss}} \tag{4.9}$$

where $\tilde{\mathbf{y}}$ is the binary prediction of the model, and (i, j) indicates the coordinate of pixels, H, W are the height and width of the image.

To convert the input color fundus image $\mathbf{x}_i$ to grayscale, I conduct the contrast limited adaptive histogram equalization (CLAHE) (Reza, 2004) on the intensity-reversed green channel and denote it as $\mathbf{x}_i^0 \in \mathscr{D}^0$. As stated before, the essential property of the generated images is that despite significant intensity variations, they consistently maintain the shared anatomical structure of the vasculature. Therefore, the $\mathscr{D}^k$ are termed anatomy-consistent pseudo-modalities. I will leverage this property in the meta-learning scheme.

4.5.1.2 Meta-learning on Anatomy Consistent Image Space

Developed from the few-shot learning paradigm, meta-learning seeks to enhance a model's generalizability to unseen data when presented with limited training sets. This is achieved by an episodic training paradigm that consists of two stages: meta-train and meta-test. The source domain $\mathscr{S}$ is split into two subsets $\mathscr{S}_{train}$ and $\mathscr{S}_{test}$ to mimic encountering OOD data during training.

Mixup is a common strategy for data augmentation as it generates new samples via linear interpolation in either image (Kim et al., 2020) or feature space (Verma et al., 2019). Zhang et al. (2020a) showed Mixup improves model generalization and robustness. In (Liu et al., 2021), the authors deploy meta-learning on generated images that are synthesized by mixing the amplitude spectrum in frequency domain. They preserve larger structures such as the optic disc by keeping the phase spectrum un-mixed. Given my anatomy-consistent pseudo-modalities, I am able to directly work on the images rather than the frequency domain. I select $\mathscr{D}^1$ as the meta-train data, and I mixup the remaining three pseudo-modalities ($\mathscr{D}^0$,$\mathscr{D}^2$, and $\mathscr{D}^3$) to form a continuous space (red facet in Figure 4.14(c)) from which I draw meta-test samples. The $\mathscr{D}^1$ is intentionally chosen to be meta-train data because it only contains the vasculature while the irrelevant features (e.g., optic disc and macular) are completely removed. Thus, it is the simplest scenario for vessel segmentation and can potentially provide a clean latent vector that only represents the vessel features. This will be further explained in the next paragraph.

In order to mixup three examples, I set a coefficient vector λ follow the Dirichlet distribution, i.e., $\lambda \sim \text{Dirichlet}(\alpha)$ where $\lambda, \alpha \in \mathbb{R}^3$. The probability density function (PDF) is defined as follows:

$$P(\lambda) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \prod_{i=1}^{3} \lambda_i^{\alpha_i - 1} \mathbb{1}(\lambda \in H) \tag{4.10}$$

with $H = \{\lambda \in \mathbb{R}^3 : \lambda_i \geq 0, \sum_{i=1}^{3} \lambda_i = 1\}$ and $\Gamma(\alpha_i) = (\alpha_i - 1)!$. Examples of PDFs with different hyperparameters α are shown in the top row of Figure 4.15.

The mixup image $\mathbf{s}_i$ is created by sampling the coefficient vector λ from $P(\lambda)$, i.e., $\mathbf{s}_i = \lambda_1 \mathbf{x}_i^0 + \lambda_2 \mathbf{x}_i^2 +$

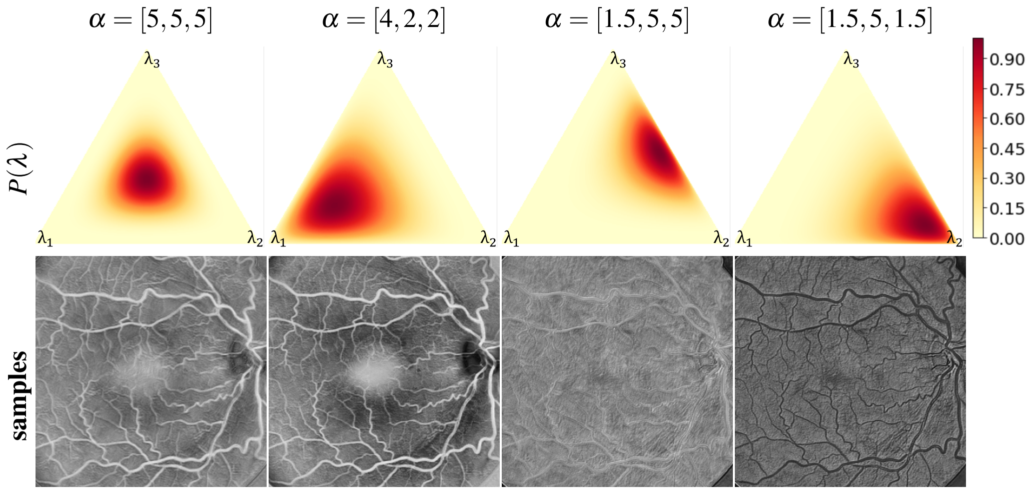


Figure 4.15: Examples of Dirichlet distribution and corresponding sample images.

$\lambda_2 \mathbf{x}_i^3$. It is evident from the bottom row of Figure 4.15 that the samples drawn from different distributions drastically vary in terms of contrast and vessel intensity. Thus, the Dirichlet mixup can augment the training data with varying styles of images without altering the vessel structure. To thoroughly exploit the continuous image space, I set $\alpha = [1,1,1]$ such that $P(\lambda)$ is a uniform distribution and all samples are considered equally. In general, each input image $\mathbf{x}_i$ representing the i^{th} vasculature anatomy is expanded into a triangular image space and an independent corner, forming a tetrahedron (Figure 4.14(c)). $\mathscr{S}_{train}$ consists of images (blue corners) in $\mathscr{D}^1$ and $\mathscr{S}_{test}$ includes mixup samples drawn from the red facet. I denote the sample images with the same vessel structure with $\mathbf{x}_i$ to be $\mathbf{s}_i^{(m)}$. The superscript with parenthesis (m) indexes the m^{th} sample.

4.5.1.3 Structural Correlation Constraints

Next, I design constraints to facilitate the model's concentration on the vessel morphology. I tackle this by delineating the correlation between latent features, as illustrated in Figure 4.14(d). For two input images $\mathbf{x}_i$ and $\mathbf{x}_j$ $(i \neq j)$, the features $\mathbf{z}_i$ and $\mathbf{z}_j$ are desired to be far apart, as their anatomies differ. In contrast, the M mixup samples $\mathbf{s}_i^{(m)}$ for $m \in \{1, \cdots, M\}$ are all anatomy-consistent, thus the corresponding features $\mathbf{z}_i^{(m)}$ should form subject-specific clusters, as shown in Figure 4.16(left). Based on this intuition, I propose two loss functions.

Similarity loss $\mathscr{L}_{sim}$. As mentioned in before, I set $\mathscr{S}_{train} = \mathscr{D}^1$. The feature vector extracted during meta-training can be regarded as an anchor in the latent space; I denote it as $\mathbf{z}_i^a$. Then the latent features $\mathbf{z}_i^{(m)}$ from samples $\mathbf{s}_i^{(m)}$, $m \in \{1, \cdots, M\}$, should be close to the anchor $\mathbf{z}_i^a$. Here, I simply use the L1 norm as the similarity loss:

$$\mathscr{L}_{sim} = \sum_{i=1}^{N} \sum_{m=1}^{M} \|\mathbf{z}_i^{(m)} - \mathbf{z}_i^a\|_1 \tag{4.11}$$

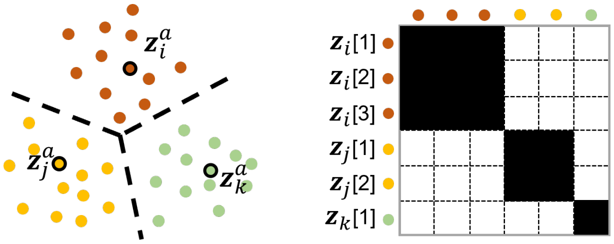


Figure 4.16: **Left**: Feature clusters. Each dot represents a feature vector. Samples representing different anatomies are shown in different colors. The highlighted dots are the latent anchor features extracted from $\mathbf{x}_i^1$, $\mathbf{x}_j^1$ and $\mathbf{x}_k^1$ during meta-training. **Right**: NCC matrix. Each entry of the matrix is the cross-correlation between two feature vectors.

where N is the number of input images, M is the number of samples. $\mathscr{L}_{sim}$ is used to reduce the distance between sample features and the anchor within the clusters, as shown in Figure 4.16(left).

Normalized cross-correlation loss $\mathscr{L}_{ncc}$. In the context of contrastive learning, the Barlow Twins objective function (Zbontar et al., 2021) was proposed to minimize the redundant information contained in the embedding vectors. This is realized by computing an empirical cross-correlation matrix of two vectors and bringing it closer to identity such that unmatched entries are not correlated. I extend this idea to a stack of vectors, as illustrated in Figure 4.16(right). Feature vectors are color coded in the same way as the left panel of the figure. The normalized cross-correlations (NCC) between each pair of features form a symmetric matrix $\mathscr{C}$. As an example, the NCC of $\mathbf{z}_i^{(3)}$ and $\mathbf{z}_j^{(2)}$:

$$\mathscr{C}_{3,5} = \mathscr{C}_{5,3} = \frac{\mathbf{z}_i^{(3)} \cdot \mathbf{z}_j^{(2)}}{\sqrt{\mathbf{z}_i^{(3)} \cdot \mathbf{z}_i^{(3)}}\sqrt{\mathbf{z}_j^{(2)} \cdot \mathbf{z}_j^{(2)}}} \tag{4.12}$$

In the ideal ground truth $\mathscr{C}^*$, the entries in the black region are 1, indicating similar features. Conversely, the white region entries are 0, representing dissimilarity. Then the NCC loss can be defined by $\mathscr{L}_{ncc} = \|\mathscr{C}^* - \mathscr{C}\|_F^2$. The total loss for the meta-test stage is:

$$\mathscr{L}_{test} = \omega_1 \mathscr{L}_{seg} + \omega_2 \mathscr{L}_{sim} + \omega_3 \mathscr{L}_{ncc} \tag{4.13}$$

Empirically, I set $\omega_1 = \omega_2 = 100$, $\omega_3 = 1$.

4.5.2 Datasets

I use 7 public datasets listed in Table 4.14. The source domain $\mathscr{S}$ includes three color fundus datasets: DRIVE, STARE and healthy samples in ARIA. By testing on the target domain $\mathscr{T}$, I evaluate the model's ability to generalize across pathological, cross-site, and cross-modality shift conditions.

dataset	modality	resolution	number	domain
DRIVE (Staal et al., 2004)	fundus	565×584	20	$\mathscr{S}$
STARE (Hoover et al., 2000)	fundus	700×605	20	$\mathscr{S}$
ARIA (Farnell et al., 2008) healthy	fundus	768×576	61	$\mathscr{S}$
AMD	fundus	768×576	59	$\mathscr{T}$
diabetic	fundus	768×576	23	$\mathscr{T}$
PRIME-FP20 (Ding et al., 2020b)	fundus	4000×4000	15	$\mathscr{T}$
ROSE (Ma et al., 2020)	OCT-A	304×304	30	$\mathscr{T}$
OCTA-500(6M) (Li et al., 2020b)	OCT-A	400×400	300	$\mathscr{T}$
RECOVERY-FA19 (Ding et al., 2020a)	FA	3900×3072	8	$\mathscr{T}$

Table 4.14: Datasets. Rows indicating the source domains have a white background while the target domains are shaded according to domain shift type. From top to bottom, (I) pathology: light gray, (II) cross-site: medium gray, (III) cross-modality: dark gray.

4.5.3 Experiments

4.5.3.1 Implementation Details

The segmentation network $g(\cdot)$ is a 6-layer residual U-Net. If the number of channels n for a layer is denoted as C_n, then the architecture is: $C_8 - C_{32} - C_{32} - C_{64} - C_{64} - C_{16}$. The synthesis model $f(\cdot)$ only functions on color fundus images in $\mathscr{S}$ during training. At test-time, fundus images are converted to grayscale by applying CLAHE on intensity-reversed green channel, while OCT-A and FA images are passed to the segmentation network $g(\cdot)$ directly. $g(\cdot)$ is trained and tested on an NVIDIA RTX 2080TI 11GB GPU. I set the batch size to 10 and train for 30 epochs. I utilize the Adam optimizer with the initial learning rate $\eta_{train} = 1 \times 10^{-3}$ for meta-training and $\eta_{test} = 5 \times 10^{-3}$ meta-testing, both decayed by 0.5 for every 3 epochs.

4.5.3.2 Competing Methods

There are three major classes of approaches to solve the DG problem: data augmentation, domain alignment, and meta-learning. I compare against a representative algorithm from each: BigAug (Zhang et al., 2020b), domain regularization network (Aslani et al., 2020), and MASF (Dou et al., 2019), respectively. I also compare to VFT (Hu et al., 2022a) as it also focuses on leveraging shape information and pseudo-modalities. Moreover, I train a residual U-Net on $\mathscr{S}$ as a baseline model, and a residual U-Net on each target domain $T^p \in \mathscr{T}$ as an oracle model, to provide an indication of the lower and upper bounds of generalization performance.

4.5.4 Results

4.5.4.1 Ablation Study

In Table 4.15, I investigate the contribution of the three major components of the proposed method: the episodic training paradigm, the similarity loss $\mathscr{L}_{sim}$ and the normalized cross-correlation loss $\mathscr{L}_{ncc}$. Note that $\mathscr{L}_{sim}$ requires the access to the latent anchor and thus is only applicable when using meta-training strategy.

Episodic	$\mathscr{L}_{sim}$	$\mathscr{L}_{ncc}$	Type I	Type II	Type III	Average
-	-	-	62.93	60.04	63.94	62.95
-	-	✓	64.73	62.48	68.06	66.02
✓	-	-	**67.50**	63.40	64.25	65.19
✓	✓	-	64.75	66.24	68.30	66.77
✓	-	✓	66.10	66.99	69.71	68.05
✓	✓	✓	67.39	**66.99**	**71.60**	**69.43**

Table 4.15: The ablation study on the main components of MAP on data with three types of distribution shift. Boldface: best result, underline: second-best result.

Method	ARIA		PRIME-FP20	OCTA 500	ROSE	RECOVERY
	amd	diabetic				
baseline	63.82	65.19	47.31	73.16	67.41	51.25
Regular	64.89	66.97	55.76	73.54	68.36	55.20
BigAug	65.55	67.27	59.97	76.88	69.32	**63.20**
MASF	65.33	67.75	65.96	77.65	67.25	50.74
VFT	61.81	64.05	54.64	77.91	72.81	48.28
MAP	**66.69**$^{\sim}$	**68.08**$^{\sim}$	**68.21**†	**78.71**†	**74.25**†	61.85†
oracle	73.34	70.65	77.80	86.57	76.03	74.54

Table 4.16: The Dice values (%) for testing on target domains. Boldface: best result, underline: second best result. $^{\sim}$: p-value ≥ 0.05, † : p-value $\ll 0.05$ in paired t-test compared to the baseline. The background is encoded the same way as Table 4.14.

Without $\mathscr{L}_{sim}$ and $\mathscr{L}_{ncc}$, the model is trained with only the segmentation loss $\mathscr{L}_{seg}$. My results show that the introduction of the episodic training provides noticeable improvement in all types of distribution shift. Both loss functions also contribute positively in general, and the proposed method ranks the best in types II and III, and second best in type I.

4.5.4.2 Comparison to Competing Methods

Table 4.16 compares the Dice coefficients (%) of the competing methods. MAP ranks the best in almost all target domains (except RECOVERY, where it ranks second), which proves that the proposed MAP algorithm effectively enhances the robustness of the model under all three domain shift conditions. For some of the datasets such as ROSE and the diabetic subset of ARIA, the MAP's performance approaches the oracle. Compared to the VFT which explicitly models the tubular vessel shape, the implicit constraints provide a better guidance for the deep model to learn the structural features.

4.5.5 Discussion

I present MAP, a method that approaches the DG problem by implicitly encouraging the model to learn about the vessel structure, which is considered to be a domain-agnostic feature. This is achieved by providing the model with synthesized images that have consistent vasculature but with significant variations in style. Then by setting constraints with regard to the correlation between latent features, the model is able to focus more on the target vessel structure. My model's generalization capability is assessed on test data with different sources of domain shift, including data with pathological phenotypes, cross-site shifts, and cross-modality shifts. The results indicate that the proposed method can greatly improve the robustness of the deep learning models across all three domain shift configurations. This work was published at the *International Conference on Medical Image Computing and Computer-Assisted Intervention (2023)*. The code is publicly available at https://github.com/DeweiHu/MAP.

4.6 AdaptDiff

In previous sections, I tried to solve the domain generalization problem under the scenario that the target data is completely unknown. However, this condition can sometimes be relaxed since it is possible that the unlabeled target data is available in training. Therefore, the unsupervised domain adaptation (UDA) is most commonly used among all the categories of DA algorithms. But, the UDA can be particularly challenging when the domain shift is substantial (e.g., in different modalities). In this section, I propose an UDA method dubbed *AdaptDiff* to approach the cross-modality retinal vessel segmentation task. While the standard modalities used to visualize the retinal vessels are OCT angiography (OCT-A) and fluorescein angiography (FA), most manual annotation of retinal vessels are conducted on fundus photography (FP) (Budai et al., 2013; Hoover et al., 2000; Farnell et al., 2008; Staal et al., 2004) as FP is easier to segment than OCT-A, and more widely available than FA. Availability of labeled public datasets in OCT-A (Li et al., 2020b; Ma et al., 2020) and FA (Ding et al., 2020a) is very limited in contrast. Thus, I used the labeled fundus photography datasets as source domains and transfer the knowledge to unlabeled OCT-A and fluorescein angiography (FA). Figure 4.17 provides an example of each dataset and shows the domain shift between the aforementioned modalities.

My general idea is to train a conditional diffusion probabilistic model (Ho et al., 2020) to synthesize target domain image from a given binary vessel mask. Similar models based upon spatially-adaptive normalization block (Park et al., 2019) has been presented in data augmentation for histology (Yu et al., 2023; Oh and Jeong, 2023) and colonoscopy (Du et al., 2023). However, all these diffusion models are trained with annotated data in the same modality. In this study, I explore, for the first time, training the conditional diffusion model in a weakly supervised manner for the cross-modality scenario.

Fig. 4.18 summarizes the primary steps of our method. Similar to Figure 4.17, I color-code the source domains $\mathscr{S}$ with blue, target domains $\mathscr{T}$ with red, and the label space as gray. First, using the labeled source domain data $\{\mathbf{x}_i^{\mathscr{S}}, \mathbf{y}_i\}_{i=1}^{|\mathscr{S}|}$, I train a segmentation model $f_{seg} : \mathbf{x}^{\mathscr{S}} \rightarrow \mathbf{y}$. Next, I utilize the f_{seg} to create pseudo-labels for target domain images $\{\mathbf{x}_j^{\mathscr{T}}, \hat{\mathbf{y}}_j\}_{j=1}^{|\mathscr{T}|}$, where $\hat{\mathbf{y}}_j = f_{seg}(\mathbf{x}_j^{\mathscr{T}})$. With the pseudo-labeled data, I train a semantic conditional diffusion model $f_{syn} : \mathbf{y} \rightarrow \mathbf{x}^{\mathscr{T}}$ that maps the vessel masks to the target domain. Experimentally, I show that even with the noisy labels, the resultant diffusion model is still sufficient to represent the target data distribution. *Since both source and target domains are images delineating human retinal vasculature, I assume that they share the same label space.* Therefore, I can generate paired target domain data by sampling with source domain labels $\hat{\mathbf{x}}^{\mathscr{T}} = f_{syn}(\mathbf{y})$. Finally, I fine-tune the segmentation network f_{seg} with the synthetic paired data $\{\hat{\mathbf{x}}^{\mathscr{T}}, \mathbf{y}\}$ to adapt to the target domain $\mathscr{T}$. I evaluate AdaptDiff on 7 public datasets (3 FP datasets for training, 2 OCT-A and 2 FA datasets for testing) and show it significantly

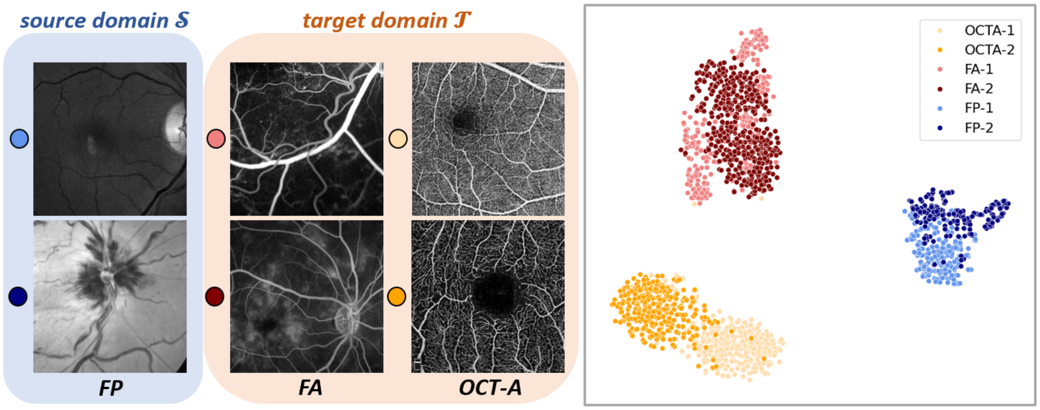
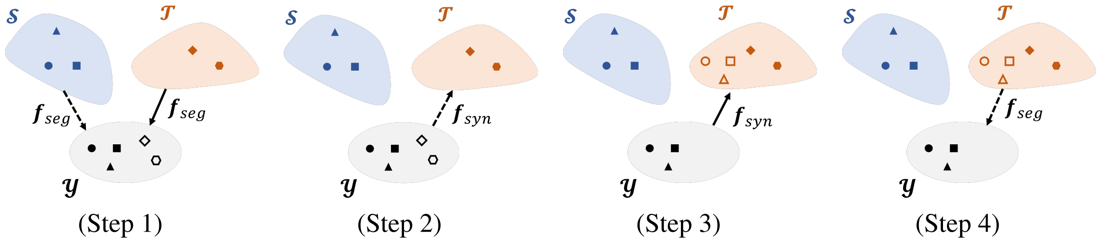


Figure 4.17: **Left**: example patches for datasets in FP, FA and OCTA. The outlines color-code the modality (red: FA, yellow: OCT-A, blue: FP). **Right**: T-SNE plot to visualize the separation between domains in feature space (extracted by pre-trained VGG-16 (Simonyan and Zisserman, 2014)).



(Step 1) (Step 2) (Step 3) (Step 4)

Figure 4.18: **(Step 1)** Train the segmentation model f_{seg} on labeled source domain and test on target domain images to create pseudo-labels. **(Step 2)** Train a semantic conditional diffusion model f_{syn} with $\{\mathbf{x}^{\mathscr{T}}, \hat{\mathbf{y}}\}$. **(Step 3)** Inference the synthetic model to generate target domain samples corresponding to the real labels $\mathbf{y}$. **(Step 4)** Fine-tune the segmentation model f_{seg} on the target domain with $\{\hat{\mathbf{x}}^{\mathscr{T}}, \mathbf{y}\}$. *Dashed/solid lines*: model training/testing. Different *marker shapes* represent distinct anatomies. *Solid* shapes are real images and manual labels, *outlines* are synthetic images and pseudo-labels.

improves the segmentation performance for cross-modality data.

4.6.1 Method

In this study, I use multiple labeled FP datasets as the source domain denoted as $\{\mathbf{x}_i^{\mathscr{S}}, \mathbf{y}_i\}_{i=1}^{|\mathscr{S}|}$ where $|\mathscr{S}|$ represents the total number of paired data. The target domains are unlabeled OCT-A or FA images denoted as $\{\mathbf{x}_j^{\mathscr{T}}\}_{j=1}^{|\mathscr{T}|}$. Since all three modalities are utilized to image the human retinal vasculature, I assume that there is no significant structural difference between underlying ground truth across different domains. In other words, **I ignore the potential distribution shift in binary vessel masks y from different datasets.**

UDA via image synthesis. Based upon the assumption above, I aim to develop a synthetic model f_{syn} that can generate a realistic target domain image conditioned on a given binary mask, i.e., $f_{syn} : \mathbf{y} \to \mathbf{x}^{\mathscr{T}}$. The generated target image is denoted as $\hat{\mathbf{x}}^{\mathscr{T}} = f_{syn}(\mathbf{y})$. Note that I can use the existing binary vessel masks $\mathbf{y}$ in the labeled source domain. In this way, I can acquire paired data $\{\hat{\mathbf{x}}^{\mathscr{T}}, \mathbf{y}\}$ on any unlabeled unseen dataset

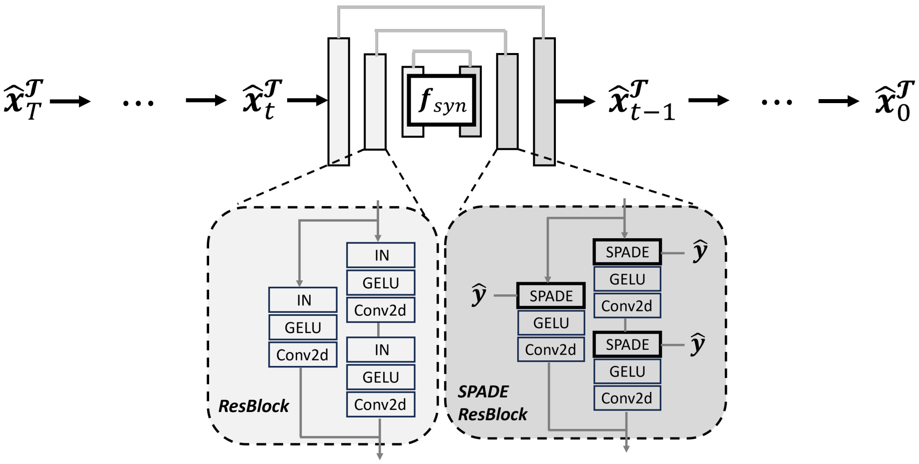


Figure 4.19: Weakly supervised conditional diffusion model. The semantic condition $\hat{\mathbf{y}}$ is added to the residual U-Net model by the spatial normalization block (SPADE) (Park et al., 2019)

.

and conduct supervised learning to get a domain adapted segmentation model. Since it has been proved that the diffusion probabilistic model is capable of superior performance in image synthesis over generative adversarial networks (GANs) (Dhariwal and Nichol, 2021), I propose to implement a semantic conditional diffusion model for f_{syn} so that the resultant image is paired with the label. However, in all the previous approaches (Yu et al., 2023; Oh and Jeong, 2023), the f_{syn} is trained with annotated data, which requires labels on target domains. In this work, I relax this constraint and demonstrate that f_{syn} can be trained in weakly supervised condition in the next paragraph. As illustrated in Figure 4.18, the proposed method takes four steps to get a domain-specific segmentation model on any unseen dataset.

Step 1: I train a segmentation model f_{seg} with the labeled source domain data $\{\mathbf{x}_i^{\mathcal{S}}, \mathbf{y}_i\}_{i=1}^{|\mathcal{S}|}$. Specifically, I pre-process the green channel of the fundus photography image with CLAHE (Reza, 2004) to enhance the contrast between the vessels and the background. Also, I invert the intensity so that the vessels are bright. Then, the model is tested on the target domain images $\{\mathbf{x}_j^{\mathcal{T}}\}_{j=1}^{|\mathcal{T}|}$ to create corresponding pseudo-labels $\{\hat{\mathbf{y}}_j\}_{j=1}^{|\mathcal{T}|}$.

Step 2: With the pseudo-labels, I am able to train the conditional diffusion model f_{syn} with $\{\mathbf{x}_j^{\mathcal{T}}, \hat{\mathbf{y}}_j\}_{j=1}^{|\mathcal{T}|}$ in a weakly supervised manner. In the following subsection, I show that the imperfect pseudo-labels $\hat{\mathbf{y}}$ and even noisier labels are sufficient for the model to learn about the conditional semantic information.

Step 3: I input the source domain labels $\{\mathbf{y}_i\}_{i=1}^{|\mathcal{S}|}$ as semantic conditional information to the diffusion model f_{syn} to generate synthetic target data $\hat{\mathbf{x}}_i^{\mathcal{T}} = f_{syn}(\mathbf{y}_i)$.

Step 4: Fine-tune the segmentation model f_{seg} on the synthetic paired data on target domain $\{\hat{\mathbf{x}}_i^{\mathcal{T}}, \mathbf{y}_i\}_{i=1}^{|\mathcal{T}|}$.

Weakly supervised conditional diffusion. For the synthetic model f_{syn}, I implement the semantic condi-

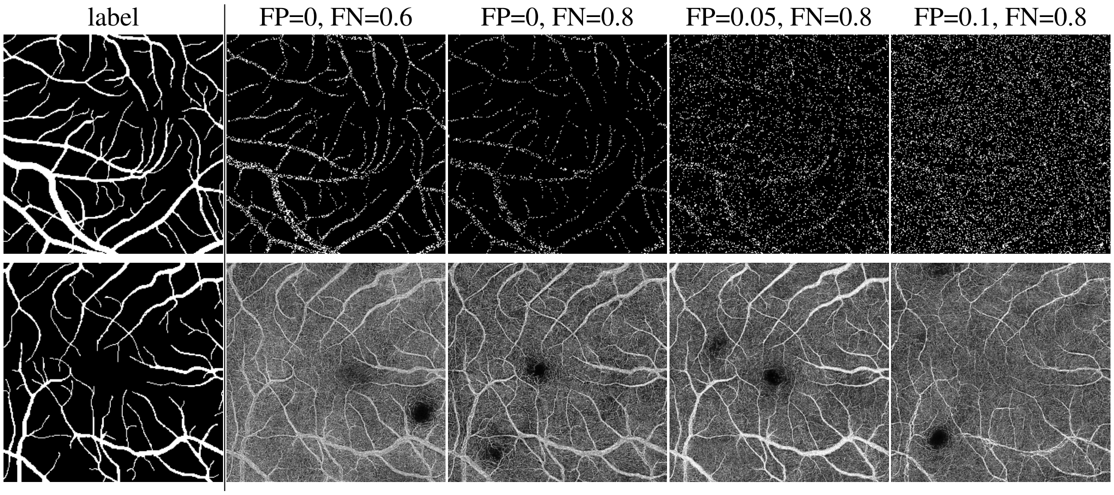


Figure 4.20: **Top**: Polluted binary masks generated from the leftmost column, used to train the semantic conditional diffusion model. FP, FN: random false positive and false negative ratios. **Bottom**: Images synthesized by models trained on corresponding masks above, given the label in the first column. Even with very high FN levels, the conditional model can generate good images.

tional diffusion model illustrated in Figure 4.19 to generate $\hat{\mathbf{x}}^{\mathscr{T}}$. Instead of using the classifier-free guidance (Wang et al., 2022), I utilize a weak supervision with pseudo-label $\hat{\mathbf{y}}$. During training, Gaussian noise is added to the real target domain image $\mathbf{x}^{\mathscr{T}}$ with respect to a time step $t \in [1,T]$ in a fixed increasing variance schedule defined by $\{\beta_1, \cdots, \beta_T\}$, where $\beta_t \in (0,1)$. Since the sum of Gaussians is still a Gaussian, the forward process is then defined by:

$$\mathbf{x}_t^{\mathscr{T}} = \sqrt{\bar{\alpha}_t}\mathbf{x}_0^{\mathscr{T}} + \sqrt{1-\bar{\alpha}_t}\varepsilon, \quad \varepsilon \sim \mathscr{N}(0,\mathbf{I}) \tag{4.14}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_t$. The synthetic model is trained to predict the noise ε of each step t in the reverse process, i.e., $\hat{\varepsilon} = \varepsilon_\theta(\hat{\mathbf{x}}_t^{\mathscr{T}}, \hat{\mathbf{y}}, t)$, where $\hat{\mathbf{y}} = f_{seg}(\mathbf{x}^{\mathscr{T}})$ is the pseudo-label and θ represents the learnable parameters in the deep learning model. The reverse process step is accomplished by reparameterization sampling:

$$\hat{\mathbf{x}}_{t-1}^{\mathscr{T}} = \frac{1}{\sqrt{\alpha_t}}\left[\hat{\mathbf{x}}_t^{\mathscr{T}} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(\hat{\mathbf{x}}_t^{\mathscr{T}}, \hat{\mathbf{y}}, t)\right] + \frac{\beta_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\delta \tag{4.15}$$

where $\delta \sim \mathscr{N}(0,\mathbf{I})$. Note that the conditional image synthesis $f_{syn}(\mathbf{y}) = \hat{\mathbf{x}}^{\mathscr{T}}$ is achieved by iteratively sampling T times from $\hat{\mathbf{x}}_T^{\mathscr{T}}$ to $\hat{\mathbf{x}}_0^{\mathscr{T}}$ using Equation 4.15.

In my experiments, I show that the semantic condition $\mathbf{y}$ does not necessarily need to be rigorously labeled by experts. Instead, pseudo-labels provided by f_{seg} trained on source domain are sufficient to help the diffusion network learn about the correlation between the semantic mask and the image. To test the robustness of the conditional diffusion model to imperfections on the labels used as condition, I create some low-quality

dataset	modality	resolution	number	domain
DRIVE Staal et al. (2004)	fundus	565×584	20	$\mathscr{S}$
STARE Hoover et al. (2000)	fundus	700×605	20	$\mathscr{S}$
HRF Budai et al. (2013)	fundus	3504×2336	45	$\mathscr{S}$
ROSE Ma et al. (2020)	OCT-A	304×304	30	$\mathscr{T}$
OCTA-500(3M) Li et al. (2020b)	OCT-A	400×400	200	$\mathscr{T}$
OCTA-500(6M) Li et al. (2020b)	OCT-A	400×400	300	$\mathscr{T}$
RECOVERY-FA19 Ding et al. (2020a)	FA	3900×3072	8	$\mathscr{T}$
PRIME Ding et al. (2020b)	FA	4000×4000	15	$\mathscr{T}$

Table 4.17: Data. Source domains have a white background; target domains are shaded.

binary vessel masks by randomly adding increasing amounts of false positives (FP) and false negatives (FN) to the original labels y from the source datasets (Fig. 4.20 top row). Then for each level of noise, I train a weakly conditional diffusion model using the OCTA500 (6M) as the target dataset. During inference, I utilize each model to generate an image conditioned on the complete binary mask from the STARE source dataset. As shown in Figure 4.20, even when trained with very noisy labels that includes 5% of FP and 80% of FN, the diffusion model can still capture all the vessels presented in testing. If the pseudo-label continues to get worse, in the rightmost scenario in Figure 4.20, many small vessels are missed in the synthesized image. In general, the proposed conditional model is very robust with regard to noisy masks. In practice, the pseudo-labels are unlikely to be as bad as the last two cases in Figure 4.20, so the f_{syn} is able to generate realistic paired data on target domains.

4.6.2 Datasets

I use 7 public datasets (Table 4.17). The source domain $\mathscr{S}$ is composed by 3 FP datasets: DRIVE, STARE and HRF. The target domains $\mathscr{T}$ contains 2 OCTA datasets, OCTA500 and ROSE, and 2 FA datasets, RECOVERY and PRIME. Note that PRIME does not have manual labels, so I use it for qualitative evaluation only.

4.6.3 Experiments

4.6.3.1 Implementation details

The diffusion model f_{syn} is a U-Net architecture with spatially adapted normalization blocks that extract semantic information from $\hat{\mathbf{y}}$ and a temporal sinusoidal encoding for time step t. For training, I set up the noise schedule of $T = 300$ with a linearly increasing variance, $\beta_0 = 1 \times 10^{-4}$ and $\beta_T = 0.02$. The model is trained for 100 epochs with mean square (MSE) loss. The initial learning rate is 1×10^{-3} and will decay by 0.5 for every 5 epochs. The segmentation network f_{seg} is a residual U-Net. The fine-tuning takes 20 epochs with Dice loss and cross-entropy loss. The initial learning rate is 5×10^{-3} and will decay by 0.5 for every 4

Method	OCTA 500(3M)	OCTA 500(6M)	ROSE	RECOVERY
baseline	65.97 ± 7.56	73.16 ± 4.54	67.41 ± 3.35	60.98 ± 4.20
DANN	69.27 ± 9.81	80.93 ± 3.31	70.37 ± 8.21	66.33 ± 3.77
CycleGAN	31.97 ± 3.90	37.47 ± 3.13	16.05 ± 2.31	22.95 ± 2.41
SynSeg	44.15 ± 3.86	47.99 ± 3.29	42.42 ± 3.09	36.68 ± 2.57
AdaptDiff	**72.73 ± 5.34**†	**81.94 ± 2.76**†	**76.15 ± 2.55**†	**71.38 ± 2.72**†
oracle	87.61 ± 2.13	84.78 ± 2.79	78.74 ± 2.47	77.13 ± 2.72

Table 4.18: Dice scores (%) for testing on target domains. Boldface: best result, underline: second best result. † : p-value $\ll 0.05$ in paired t-test compared to the baseline.

epochs. Both models are trained and tested on an NVIDIA RTX 2080TI 11GB GPU.

4.6.3.2 Competing methods

As discussed in Section 4.1, the major previous works about deep UDA can be categorized into (1) feature alignment and (2) image alignment. I select one representative work from each category as competing methods to illustrate the effectiveness of AdaptDiff. For feature alignment, Ganin and Lempitsky (2015) introduce the domain-adversarial neural network (DANN) which utilizes a gradient reversal layer is implemented in an adversarial network framework to enforce the comparability of feature map distributions across domains. Based on DANN, Javanmardi and Tasdizen (2018) propose a domain adaptation method on FP vessel segmentation. Here, I include their method as one of the competing methods. As for the image alignment, I implement the CycleGAN (Zhu et al., 2017) and SynSeg (Huo et al., 2018) for comparison.

4.6.4 Results

4.6.4.1 Quantitative results

In Table 4.18, I evaluate the UDA segmentation performance by the Dice score. Since the proposed AdaptDiff can generate realistic target domain images that closely match the label, its adaptation performance is the best in all target datasets. For OCTA500 (6M) and ROSE, my results are very close to the oracle model which is directly trained on the target domain. DANN has a reasonable performance above the baseline, which directly uses the f_{seg} without adaptation. I note that CycleGAN and SynSeg have even lower performance than the baseline. To understand this behavior, I next look at qualitative results.

4.6.4.2 Qualitative results

In Figure 4.21 I compare the quality of the synthesized images on different target domains generated by the three image alignment approaches. I observe that the fake images generated by CycleGAN and SynSeg are not well correlated with the label, even though both models are trained to map from source domain image $\mathbf{x}^{\mathscr{S}}$ to target domain image $\mathbf{x}^{\mathscr{T}}$. This is likely because there are strong domain-specific features in both

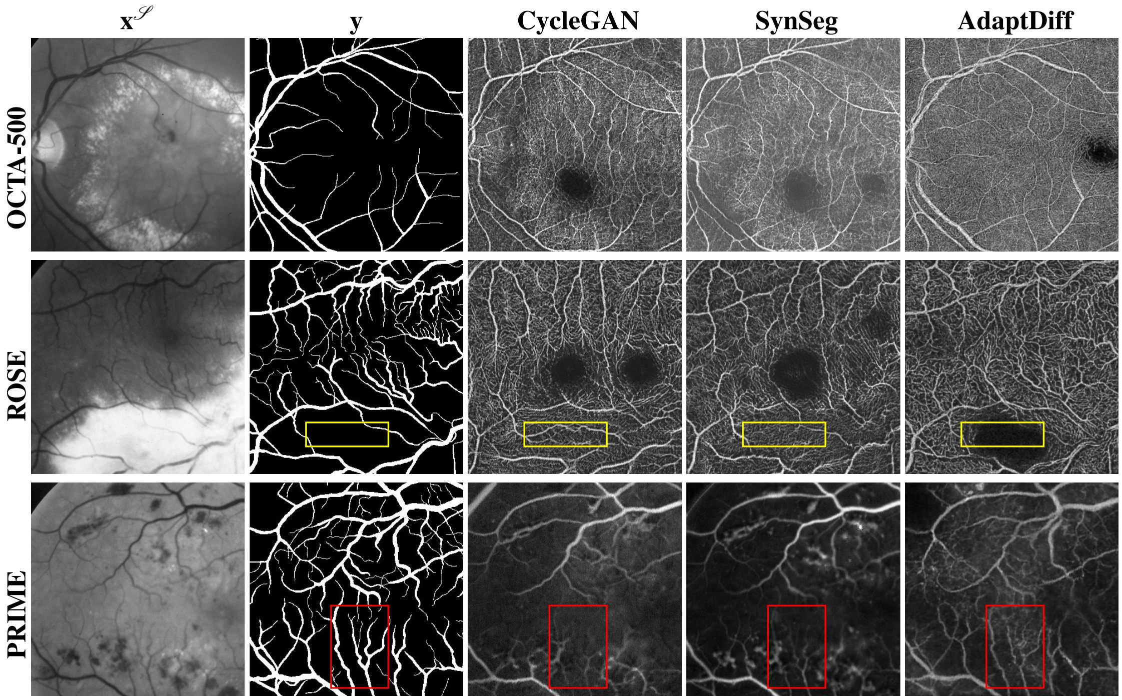


Figure 4.21: Synthetic images from the image alignment approaches. The yellow boxes highlight spurious vessels, red boxes highlight missing vessels, for CycleGAN and SynSeg outputs. AdaptDiff-synthesized images have the correct anatomy in both areas.

modalities, such as the pathological features in FP and the foveal avascular zone (FAZ) in OCT-A. Moreover, unlike brain and abdomen CT-MR translation, retinal images have a wide variability of vessel anatomy that falls within the image field of view. Consequently, it is hard to learn the mapping function between these modalities with certain structural features preserved. In contrast, AdaptDiff successfully synthesizes images that match the provided vessel labels.

4.6.5 Discussion

In this section, I proposed AdaptDiff, a diffusion-based solution to the cross-modality UDA problem. I empirically showed that the conditional semantic diffusion model can be trained in a weakly supervised manner even with rather poor pseudo-labels. Thus, for each unseen domain, I can generate paired data to fine-tune the segmentation model to significantly alleviate the effects of domain shift. The AdaptDiff can potentially be extended to a source-free domain adaptation method since it can tolerate severe noise in the label. A straightforward way is to get pseudo-labels with a model-based algorithm e.g., traditional Frangi (Frangi et al., 1998) filter and Otsu thresholding (Otsu, 1979) to acquire low quality pseudo-labels. Then we can directly train the segmentation network on target domains. Other than this, as I don't consider the distribution shift in the binary vessel masks in this study, accounting for any domain shifts in the label space remains as future work.

4.7 Chapter summary

In this chapter, I thoroughly investigate the domain generalization and domain adaptation methods that can potentially improve the robustness of the deep learning model on unseen target domain data. This is essential for medical image analysis area because the annotated data is always limited and the domain shift between different datasets can be significant. Specifically for the retinal vessel segmentation task, manually labeling vessels on OCT-A and FA is a challenging task, and thus there are very few annotated public datasets in these two modalities. In contrast, the fundus photography is relatively easier to work on, and there are sufficient labeled data that I can get access to. Given that my target objects, the retinal vessels, have a domain-invariant tubular shape, I propose both explicit (Section 4.3, Section 4.4) and implicit (Section 4.5) approaches to model it. As the results suggest, the presented DG methods can effectively improve the generalizability of the vessel segmentation model even with small amount of training data. Particularly, the implicit representation of the tubular shape has advantages in processing time compared with the explicit counterpart since it does not require the computation of Hessian matrices beforehand. More importantly, it has better performance in terms of Dice coefficient. In Section 4.6, I further explore the possibility of training the model with synthesized images. The proposed AdaptDiff further relaxes the need of manual annotation for retinal vessels. In conclusion, my methods are able to boost the deployment of learning-based algorithms in real medical application.

CHAPTER 5

Future work and Conclusion

Aim 1: OCT image denoising. In Chapter 2, I introduce both a weakly supervised and an unsupervised method for denoising OCT B-scans. The self-fusion technique, as described by Oguz et al. (2020), serves as an additional supervision source to impose the prior knowledge of the structural consistency between adjacent b-scans during the training phase. This approach effectively reduces speckle noise while preserving fine details such as retinal vessels. However, quantitative evaluation of denoising performance remains challenging due to the absence of noise-free reference images and the limitations of existing metrics in accurately reflecting the quality of denoised results. Another issue is the scarcity of public datasets and the variability in speckle patterns across different imaging systems, which typically restricts the testing of OCT denoising algorithms to in-house datasets. This lack of standardized benchmarks complicates the comparison of algorithmic performance. Future research should therefore focus on developing more robust evaluation metrics and establishing a comprehensive benchmark for this task. Although the paired noise-free reference b-scans are not likely to be accessible, there are some blind image quality assessment metrics to validate the naturalness of the results. For example, the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) (Mittal et al., 2012a), Naturalness Image Quality Evaluator (NIQE) (Mittal et al., 2012b), and Perception-based Image Quality Evaluator (PIQUE) (Venkatanath et al., 2015) are employed to quantitatively assess the level of naturalness in the images on synthetic datasets. These approaches are based on the measurement of the deviations from statistical regularities observed in normal natural images. In other words, a well-defined clean image distribution should be constructed. Given sufficient clean OCT b-scans images, similar blind assessments can be conducted without paired ground truth.

Aim 2: Retinal vessel enhancement. In Chapter 3, I introduce a novel synthesis network designed to generate enhanced 2D angiograms based on shared feature extraction. Leveraging the inherent randomness in stochastic gradient descent, I successfully generate three stable pseudo-modalities that are later utilized as augmented data in a domain generalization project. This approach is further extended to unsupervised 3D vessel segmentation on OCT-A. For experimental validation, I manually labeled the OCT-A of a small region in the human retina and a whole volume of zebrafish retina, though this limited dataset is insufficient for definitive conclusions. Given the challenges associated with manually annotating real volumetric data, there is potential to simulate 3D binary vasculature and generate realistic data based on this simulated ground truth, as suggested by Kreitner et al. (2024). In the following up research, the 3D segmentation model can be trained and tested on these generated paired data, enhancing both the model's robustness and applicability.

Aim 3: Domain adaptation and generalization. In Chapter 4, I explore domain generalization (DG) and adaptation (DA) methods, enabling the deep segmentation model to be trained on fundus photography data and tested on OCT-A and fluorescein angiography. For DG, I captured the domain-invariant morphological structure of the vessels using both explicit and implicit representations, which significantly enhanced the model's generalizability. In the DA approach, I employed a semantic conditional diffusion probabilistic model to generate realistic training data for the target domain. This model has proven effective even with noisy labels. The findings suggest potential for expanding this work into a source-free strategy using unlabeled target data, further broadening its applicability and effectiveness in real-world scenarios. Instead of training a segmentation network on the source domain, the pseudo-label can be acquired in an unsupervised manner by simply conducting the Frangi filter (Frangi et al., 1998) and Otsu thresholding (Otsu, 1979). In this way, there is no source domain needed, the weak conditional diffusion model can be directly trained on any unlabeled datasets. The only material required is the real binary vessel mask for synthetic paired data generation. Kreitner et al. (2024) recently propose an algorithm to simulate the 3D retinal vasculature. They set up a graph growing process governed by the angiogenesis in biological organisms. Different settings of hyper-parameters can result in divergence in diameter and density of the binary vessel masks. Therefore, the source-free domain adaptation method can potentially get rid of any manual labels and be fully trained with generated paired data. Moreover, such label simulator enables the discussion about the distribution shift between the ground truth annotated on different datasets. In all my works, I assume that there is no domain shift in label space due to the lack of resources to discern such a shift. However, this assumption does not hold in many scenarios in practice. For example, Li et al. (2020b) recently posted an update on the OCTA-500 dataset. They provide a more detailed annotation that includes all the small capillaries that are ignored in the previous version. Apparently, for researchers who are interested in the deep vascular complex (DVC), the segmentation of capillaries is essential, and a model trained with superficial vascular complex (SVC) labels is not helpful. Such difference in annotations prevails across domains. With the simulation method, the divergence in manual annotation can be modeled with the settings of hyper-parameters.

Another follow-up research direction is to impose the semantic condition to the diffusion model at test time. Basically, the diffusion model is directly trained with unlabeled target data without any condition. Then at test time, the anatomical structure is determined during the sampling process. In this way, the acquisition of the pseudo-label is omitted. Dhariwal and Nichol (2021) propose an idea to implement the gradient of a loss function to push the sample mean in each denoising step so that it converges in a specific region in the target image distribution. Experimentally, I observe that the strength of the gradient needs to be large and the resultant image suffers from serious hallucination problems. Therefore, an additional source of constraint is needed to improve the quality of the generated images.

As for the DG, the future research will focus more on the foundation model. The Segment Anything Model (SAM) (Kirillov et al., 2023) has been widely applied in many medical image segmentation tasks (Li et al., 2024; Yao et al., 2024). In recent research, Wang et al. (2024) explore the SAM application on OCT-A vessel segmentation. Since the retinal vasculature is composed with many disconnected plexuses, each individual connected component may need a positive prompt which make the method very tedious. Therefore, it requires more endeavor to solve this problem.

Conclusion. In my research, I focus on reducing the dependency on manual annotation for deep learning solutions on OCT image denoising and retinal vessel segmentation. One essential idea is to impose specific constraints to the deep learning model in addition to its data-driven mechanism. Such control is usually based upon traditional model-based algorithms such as self-fusion (Oguz et al., 2020) and Frangi filter (Frangi et al., 1998).

Although deep learning methods are able to delineate very complex non-linearity and have accomplished great performance in many tasks, their underlying reasoning is a black box that needs to be regularized to satisfy the specific requirements of the task. The conventional handcrafted filters have remarkable advantages in reducing the need for labels and bridging the gap between domains. In general, I combine the classic algorithms with deep learning and get promising results for unsupervised OCT image denoising and domain generalization.

Another important aspect in my research is the utilization of the synthesized images as an auxiliary resource to train the model. Despite general skepticism about the validity of synthesized images, and the evaluation of their fidelity is a non-trivial task, they can still contribute to the training as long as they are not used as direct output. Therefore, I implement the pseudo-modalities generated from fundus photography images as the augmented data in the domain generalization project.

These two broad contributions are general methodologies that can be extended to the application of deep learning methods to many other tasks and modalities in the field of medical image analysis.

References

Abouzeid, H. and Wolfensberger, T. J. (2006). Macular recovery after retinal detachment. *Acta ophthalmologica Scandinavica*, 84(5):597–605.

Adhi, M. and Duker, J. S. (2013). Optical coherence tomography–current and future applications. *Current opinion in ophthalmology*, 24(3):213.

Adiga, S. and Sivaswamy, J. (2018). Shared encoder based denoising of optical coherence tomography images. In *ICVGIP*, pages 35–1.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Aslani, S., Murino, V., Dayan, M., Tam, R., Sona, D., and Hamarneh, G. (2020). Scanner invariant multiple sclerosis lesion segmentation from mri. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 781–785. IEEE.

Avanaki, M. R., Cernat, R., Tadrous, P. J., Tatla, T., Podoleanu, A. G., and Hojjatoleslami, S. A. (2013). Spatial compounding algorithm for speckle reduction of dynamic focus oct images. *IEEE Photonics Technology Letters*, 25(15):1439–1442.

Bao, P. and Zhang, L. (2003). Noise reduction for magnetic resonance images via adaptive multiscale products thresholding. *IEEE transactions on medical imaging*, 22(9):1089–1099.

Bek, T. (2017). Diameter changes of retinal vessels in diabetic retinopathy. *Current diabetes reports*, 17:1–7.

Bock, R., Meier, J., Michelson, G., Nyúl, L. G., and Hornegger, J. (2007). Classifying glaucoma with image-based features from fundus photographs. In *Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29*, pages 355–364. Springer.

Bowd, C., Weinreb, R. N., Williams, J. M., and Zangwill, L. M. (2000). The retinal nerve fiber layer thickness in ocular hypertensive, normal, and glaucomatous eyes with optical coherence tomography. *Archives of ophthalmology*, 118(1):22–26.

Budai, A., Bock, R., Maier, A., Hornegger, J., and Michelson, G. (2013). Robust vessel segmentation in fundus images. *International journal of biomedical imaging*, 2013.

Burke, T. R., Chu, C. J., Salvatore, S., Bailey, C., Dick, A. D., Lee, R. W., Ross, A. H., and Carreño, E. (2017). Application of oct-angiography to characterise the evolution of chorioretinal lesions in acute posterior multifocal placoid pigment epitheliopathy. *Eye*, 31(10):1399–1408.

Cabrera DeBuc, D., Somfai, G. M., and Koller, A. (2017). Retinal microvascular network alterations: potential biomarkers of cerebrovascular and neural diseases. *American Journal of Physiology-Heart and Circulatory Physiology*, 312(2):H201–H212.

Cai, N., Shi, F., Gu, Y., Hu, D., Chen, Y., and Chen, X. (2018). A resnet-based universal method for speckle reduction in optical coherence tomography images. In *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*.

Canero, C. and Radeva, P. (2003). Vesselness enhancement diffusion. *Pattern Recognition Letters*, 24(16):3141–3151.

Chen, C., Bai, W., Davies, R. H., Bhuva, A. N., Manisty, C. H., Augusto, J. B., Moon, J. C., Aung, N., Lee, A. M., Sanghvi, M. M., et al. (2020a). Improving the generalizability of convolutional neural network-based segmentation on cmr images. *Frontiers in cardiovascular medicine*, 7:105.

Chen, I.-L., Ho, T.-S., and Lu, C.-W. (2020b). Full field optical coherence tomography image denoising using deep learning with spatial compounding. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1975–1978. IEEE.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, Q., Yu, X., Sun, Z., Dai, H., et al. (2016). The application of octa in assessment of anti-vegf therapy for idiopathic choroidal neovascularization. *Journal of ophthalmology*, 2016.

Chen, Z., Zeng, Z., Shen, H., Zheng, X., Dai, P., and Ouyang, P. (2020c). Dn-gan: Denoising generative adversarial networks for speckle noise reduction in optical coherence tomography images. *Biomedical Signal Processing and Control*, 55:101632.

Cheung, C. Y.-l., Ikram, M. K., Chen, C., and Wong, T. Y. (2017). Imaging retina to study dementia and stroke. *Progress in retinal and eye research*, 57:89–107.

Chiu, S. J., Allingham, M. J., Mettu, P. S., Cousins, S. W., Izatt, J. A., and Farsiu, S. (2015). Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. *Biomedical optics express*, 6(4):1172–1194.

Chui, T. Y., Zhong, Z., Song, H., and Burns, S. A. (2012). Foveal avascular zone and its relationship to foveal pit shape. *Optometry and Vision Science*, 89(5):602–610.

Cincotti, G., Loi, G., and Pappalardo, M. (2001). Frequency decomposition and compounding of ultrasound medical images with wavelet packets. *IEEE transactions on medical imaging*, 20(8):764–771.

Cogan, D. G., Toussaint, D., and Kuwabara, T. (1961). Retinal vascular patterns: Iv. diabetic retinopathy. *Archives of ophthalmology*, 66(3):366–378.

Devalla, S. K., Subramanian, G., Pham, T. H., Wang, X., Perera, S., Tun, T. A., Aung, T., Schmetterer, L., Thiéry, A. H., and Girard, M. J. (2019). A deep learning approach to denoise optical coherence tomography images of the optic nerve head. *Scientific reports*, 9(1):1–13.

Dewey, B. E., Zhao, C., Reinhold, J. C., Carass, A., Fitzgerald, K. C., Sotirchos, E. S., Saidha, S., Oh, J., Pham, D. L., Calabresi, P. A., et al. (2019). Deepharmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic resonance imaging*, 64:160–170.

Dewey, B. E., Zuo, L., Carass, A., He, Y., Liu, Y., Mowry, E. M., Newsome, S., Oh, J., Calabresi, P. A., and Prince, J. L. (2020). A disentangled latent space for cross-site mri harmonization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 720–729. Springer.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*.

Ding, L., Bawany, M. H., Kuriyan, A. E., Ramchandran, R. S., Wykoff, C. C., and Sharma, G. (2020a). A novel deep learning pipeline for retinal vessel detection in fluorescein angiography. *IEEE Transactions on Image Processing*, 29:6561–6573.

Ding, L., Kuriyan, A. E., Ramchandran, R. S., Wykoff, C. C., and Sharma, G. (2020b). Weakly-supervised vessel detection in ultra-widefield fundus photography via iterative multi-modal registration and learning. *IEEE Transactions on Medical Imaging*, 40(10):2748–2758.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dou, Q., Coelho de Castro, D., Kamnitsas, K., and Glocker, B. (2019). Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32.

Du, Y., Jiang, Y., Tan, S., Wu, X., Dou, Q., Li, Z., Li, G., and Wan, X. (2023). Arsdm: colonoscopy images synthesis with adaptive refinement semantic diffusion models. In *International conference on medical image computing and computer-assisted intervention*, pages 339–349. Springer.

El-Haddad, M. T., Bozic, I., and Tao, Y. K. (2018). Spectrally encoded coherence tomography and reflectometry: Simultaneous en face and cross-sectional imaging at 2 gigapixels per second. *Journal of biophotonics*, 11(4):e201700268.

Fan, W., Yu, H., Chen, T., and Ji, S. (2020). Oct image restoration using non-local deep image prior. *Electronics*, 9(5):784.

Farnell, D., Hatfield, F., Knox, P., Reakes, M., Spencer, S., Parry, D., and Harding, S. P. (2008). Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators. *Journal of the Franklin institute*, 345(7):748–765.

Farsiu, S., Chiu, S. J., O'Connell, R. V., Folgar, F. A., Yuan, E., Izatt, J. A., Toth, C. A., Group, A.-R. E. D. S. . A. S. D. O. C. T. S., et al. (2014). Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology*, 121(1):162–172.

Fercher, A. F., Drexler, W., Hitzenberger, C. K., and Lasser, T. (2003). Optical coherence tomography-principles and applications. *Reports on progress in physics*, 66(2):239.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Frangi, A. F., Niessen, W. J., Vincken, K. L., and Viergever, M. A. (1998). Multiscale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI'98: First International Conference Cambridge, MA, USA, October 11–13, 1998 Proceedings 1*, pages 130–137. Springer.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Giarratano, Y., Bianchi, E., Gray, C., Morris, A., MacGillivray, T., Dhillon, B., and Bernabeu, M. O. (2019). Automated and network structure preserving segmentation of optical coherence tomography angiograms. *arXiv preprint arXiv:1912.09978*.

Gisbert, G., Dey, N., Ishikawa, H., Schuman, J., Fishbaugh, J., and Gerig, G. (2020). Self-supervised denoising via diffeomorphic template estimation: Application to optical coherence tomography. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 72–82. Springer.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gour, N. and Khanna, P. (2019). Speckle denoising in optical coherence tomography images using residual deep convolutional neural network. *Multimedia Tools and Applications*, pages 1–17.

Guan, H. and Liu, M. (2021). Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185.

Guizar-Sicairos, M., Thurman, S. T., and Fienup, J. R. (2008). Efficient subpixel image registration algorithms. *Optics letters*, 33(2):156–158.

Guo, A., Fang, L., Qi, M., and Li, S. (2020). Unsupervised denoising of optical coherence tomography images with nonlocal-generative adversarial network. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12.

Guo, Y., Wang, K., Yang, S., Wang, Y., Gao, P., Xie, G., Lv, C., and Lv, B. (2019). Structure-aware noise reduction generative adversarial network for optical coherence tomography image. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 9–17. Springer.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., and Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584.

Hayreh, S. S., Podhajsky, P. A., and Zimmerman, M. B. (2009). Retinal artery occlusion: associated systemic and ophthalmic abnormalities. *Ophthalmology*, 116(10):1928–1936.

Hellström, A., Smith, L. E., and Dammann, O. (2013). Retinopathy of prematurity. *The lancet*, 382(9902):1445–1457.

Hesamian, M. H., Jia, W., He, X., and Kennedy, P. (2019). Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32:582–596.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*.

Holló, G. (2018). Comparison of peripapillary oct angiography vessel density and retinal nerve fiber layer thickness measurements for their ability to detect progression in glaucoma. *Journal of glaucoma*, 27(3):302–305.

Hoover, A., Kouznetsova, V., and Goldbaum, M. (2000). Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE TMI*, 19(3):203–210.

Hore, A. and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE.

Hu, D., Cui, C., Li, H., Larson, K. E., Tao, Y. K., and Oguz, I. (2021a). Life: a generalizable autodidactic pipeline for 3d oct-a vessel segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 514–524. Springer.

Hu, D., Li, H., Liu, H., and Oguz, I. (2021b). Domain generalization for retinal vessel segmentation with vector field transformer. In *MIDL*.

Hu, D., Li, H., Liu, H., and Oguz, I. (2022a). Domain generalization for retinal vessel segmentation with vector field transformer. In *International Conference on Medical Imaging with Deep Learning*, pages 552–564. PMLR.

Hu, D., Malone, J. D., Atay, Y., Tao, Y. K., and Oguz, I. (2020). Retinal oct denoising with pseudo-multimodal fusion network. In *Ophthalmic Medical Image Analysis: 7th International Workshop, OMIA 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 7*, pages 125–135. Springer.

Hu, D., Tao, Y. K., and Oguz, I. (2022b). Unsupervised denoising of retinal oct with diffusion probabilistic model. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 25–34. SPIE.

Huang, D., Swanson, E. A., Lin, C. P., Schuman, J. S., Stinson, W. G., Chang, W., Hee, M. R., Flotte, T., Gregory, K., Puliafito, C. A., et al. (1991). Optical coherence tomography. *science*, 254(5035):1178–1181.

Huang, Y., Lu, Z., Shao, Z., Ran, M., Zhou, J., Fang, L., and Zhang, Y. (2019). Simultaneous denoising and super-resolution of optical coherence tomography images based on generative adversarial network. *Optics express*, 27(9):12289–12307.

Huang, Y., Xia, W., Lu, Z., Liu, Y., Zhou, J., Fang, L., and Zhang, Y. (2020). Disentanglement network for unsupervised speckle reduction of optical coherence tomography images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 675–684. Springer.

Huang, Y., Zhang, N., and Hao, Q. (2021). Real-time noise reduction based on ground truth free deep learning for optical coherence tomography. *Biomedical Optics Express*, 12(4):2027–2040.

Huo, Y., Xu, Z., Moon, H., Bao, S., Assad, A., Moyo, T. K., Savona, M. R., Abramson, R. G., and Landman, B. A. (2018). Synseg-net: Synthetic segmentation without target modality ground truth. *IEEE transactions on medical imaging*, 38(4):1016–1025.

Ikram, M., De Jong, F., Bos, M., Vingerling, J., Hofman, A., Koudstaal, P., De Jong, P., and Breteler, M. (2006). Retinal vessel diameters and risk of stroke: the rotterdam study. *Neurology*, 66(9):1339–1343.

Irshad, S. and Akram, M. U. (2014). Classification of retinal vessels into arteries and veins for detection of hypertensive retinopathy. In *2014 Cairo International Biomedical Engineering Conference (CIBEC)*, pages 133–136. IEEE.

Ishibazawa, A., Nagaoka, T., Takahashi, A., Omae, T., Tani, T., Sogawa, K., Yokota, H., and Yoshida, A. (2015). OCT angiography in diabetic retinopathy: a prospective pilot study. *American journal of ophthalmology*, 160(1):35–44.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Javanmardi, M. and Tasdizen, T. (2018). Domain adaptation for biomedical image segmentation using adversarial training. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 554–558. IEEE.

Jerman, T., Pernuš, F., Likar, B., and Špiclin, Ž. (2015). Beyond frangi: an improved multiscale vesselness filter. In *Medical Imaging 2015: Image Processing*, volume 9413, pages 623–633. SPIE.

Kaur, A., Singh, Y., Neeru, N., Kaur, L., and Singh, A. (2021). A survey on deep learning approaches to medical images and a systematic look up into real-time object detection. *Archives of Computational Methods in Engineering*, pages 1–41.

Keane, P. A., Patel, P. J., Liakopoulos, S., Heussen, F. M., Sadda, S. R., and Tufail, A. (2012). Evaluation of age-related macular degeneration with optical coherence tomography. *Survey of ophthalmology*, 57(5):389–414.

Keith, N. M., WAGENER, H. P., and BARKER, N. W. (1974). Some different types of essential hypertension: their course and prognosis. *The American journal of the medical sciences*, 268(6):336–345.

Khandelwal, P. and Yushkevich, P. (2020). Domain generalizer: A few-shot meta learning framework for domain generalization in medical imaging. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 73–84. Springer.

Kim, J.-H., Choo, W., and Song, H. O. (2020). Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pages 5275–5285. PMLR.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Kreitner, L., Paetzold, J. C., Rauch, N., Chen, C., Hagag, A. M., Fayed, A. E., Sivaprasad, S., Rausch, S., Weichsel, J., Menze, B. H., et al. (2024). Synthetic optical coherence tomography angiographs for detailed retinal vessel segmentation without human annotations. *IEEE Transactions on Medical Imaging*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kumari, S. and Singh, P. (2023). Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. *Computers in Biology and Medicine*, page 107912.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Law, M. W. and Chung, A. C. (2008). Three dimensional curvilinear structure detection using optimally oriented flux. In *European conference on computer vision*, pages 368–382. Springer.

Le Bihan, D., Mangin, J.-F., Poupon, C., Clark, C. A., Pappata, S., Molko, N., and Chabriat, H. (2001). Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging*, 13(4):534–546.

Lee, S. C., Tran, S., Amin, A., Morse, L. S., Moshiri, A., Park, S. S., and Yiu, G. (2020). Retinal vessel density in exudative and nonexudative age-related macular degeneration on optical coherence tomography angiography. *American journal of ophthalmology*, 212:7–16.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T. (2018). Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*.

Li, H., Hu, D., Liu, H., Wang, J., and Oguz, I. (2022). Cats: Complementary cnn and transformer encoders for segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Li, H., Liu, H., Hu, D., Wang, J., and Oguz, I. (2024). Prism: A promptable and robust interactive segmentation model with visual prompts. *arXiv preprint arXiv:2404.15028*.

Li, H., Wang, Y., Wan, R., Wang, S., Li, T.-Q., and Kot, A. (2020a). Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in Neural Information Processing Systems*, 33:3118–3129.

Li, M., Chen, Y., Ji, Z., Xie, K., Yuan, S., Chen, Q., and Li, S. (2020b). Image projection network: 3d to 2d image segmentation in octa images. *IEEE Transactions on Medical Imaging*, 39(11):3343–3354.

Liang, J., He, R., and Tan, T. (2023). A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361*.

Liesenfeld, B., Kohner, E., Piehlmeier, W., Kluthe, S., Aldington, S., Porta, M., Bek, T., Obermaier, M., Mayer, H., Mann, G., et al. (2000). A telemedical approach to the screening of diabetic retinopathy: digital fundus photography. *Diabetes care*, 23(3):345–348.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Liu, Q., Chen, C., Qin, J., Dou, Q., and Heng, P.-A. (2021). Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023.

Liu, Y., Carass, A., Zuo, L., He, Y., Han, S., Gregori, L., Murray, S., Mishra, R., Lei, J., Calabresi, P. A., et al. (2022). Disentangled representation learning for octa vessel segmentation with limited training data. *IEEE transactions on medical imaging*, 41(12):3686–3698.

Liu, Y., Zuo, L., Carass, A., He, Y., Filippatou, A., Solomon, S. D., Saidha, S., Calabresi, P. A., and Prince, J. L. (2020). Variational intensity cross channel encoder for unsupervised vessel segmentation on oct angiography. In *SPIE Medical Imaging 2020: Image Processing*, volume 11313, page 113130Y.

Lyu, J., Zhang, Y., Huang, Y., Lin, L., Cheng, P., and Tang, X. (2022). Aadg: Automatic augmentation for domain generalization on retinal image segmentation. *IEEE Transactions on Medical Imaging*.

Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., and Martel, A. L. (2021). Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035.

Ma, Y., Chen, X., Zhu, W., Cheng, X., Xiang, D., and Shi, F. (2018). Speckle noise reduction in optical coherence tomography images based on edge-sensitive cgan. *Biomedical optics express*, 9(11):5129–5146.

Ma, Y., Hao, H., Xie, J., Fu, H., Zhang, J., Yang, J., Wang, Z., Liu, J., Zheng, Y., and Zhao, Y. (2020). Rose: a retinal oct-angiography vessel segmentation dataset and new model. *IEEE transactions on medical imaging*, 40(3):928–939.

Malone, J. D., El-Haddad, M. T., Yerramreddy, S. S., Oguz, I., and Tao, Y. K. (2019). Handheld spectrally encoded coherence tomography and reflectometry for motion-corrected ophthalmic optical coherence tomography and optical coherence tomography angiography. *Neurophotonics*, 6(4):041102–041102.

Mao, Z., Miki, A., Mei, S., Dong, Y., Maruyama, K., Kawasaki, R., Usui, S., Matsushita, K., Nishida, K., and Chan, K. (2019). Deep learning based noise reduction method for automatic 3d segmentation of the anterior of lamina cribrosa in optical coherence tomography volumetric scans. *Biomedical optics express*, 10(11):5832–5851.

Medeiros, F. A., Zangwill, L. M., Bowd, C., Vessani, R. M., Susanna Jr, R., and Weinreb, R. N. (2005). Evaluation of retinal nerve fiber layer, optic nerve head, and macular thickness measurements for glaucoma detection using optical coherence tomography. *American journal of ophthalmology*, 139(1):44–55.

Midena, E., Frizziero, L., Torresin, T., Boscolo Todaro, P., Miglionico, G., and Pilotto, E. (2020). Optical coherence tomography and color fundus photography in the screening of age-related macular degeneration: A comparative, population-based study. *Plos one*, 15(8):e0237352.

Mirshahi, A., Feltgen, N., Hansen, L. L., and Hattenbach, L.-O. (2008). Retinal vascular occlusions: an interdisciplinary challenge. *Deutsches Ärzteblatt International*, 105(26):474.

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012a). No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708.

Mittal, A., Soundararajan, R., and Bovik, A. C. (2012b). Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212.

Moss, H. E. (2015). Retinal vascular changes are a marker for cerebral vascular diseases. *Current neurology and neuroscience reports*, 15:1–9.

Nguyen, T. T. and Wong, T. Y. (2009). Retinal vascular changes and diabetic retinopathy. *Current diabetes reports*, 9(4):277–283.

Oguz, I., Malone, J. D., Atay, Y., and Tao, Y. K. (2020). Self-fusion for OCT noise reduction. In *SPIE Medical Imaging 2020: Image Processing*, volume 11313, page 113130C.

Oh, H.-J. and Jeong, W.-K. (2023). Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets. *arXiv preprint arXiv:2306.14132*.

Otálora, S., Atzori, M., Andrearczyk, V., Khan, A., and Müller, H. (2019). Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. *Frontiers in bioengineering and biotechnology*, page 198.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66.

Ouyang, J., Adeli, E., Pohl, K. M., Zhao, Q., and Zaharchuk, G. (2021). Representation disentanglement for multi-modal brain mri analysis. In *International Conference on Information Processing in Medical Imaging*, pages 321–333. Springer.

Palladino, J. A., Slezak, D. F., and Ferrante, E. (2020). Unsupervised domain adaptation via cyclegan for white matter hyperintensity segmentation in multicenter mr images. In *16th International Symposium on Medical Information Processing and Analysis*, volume 11583, page 1158302. SPIE.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346.

Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. on pattern analysis and machine intelligence*, 12(7):629–639.

Qin, X., Song, X., and Jiang, S. (2023). Bi-level meta-learning for few-shot domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15900–15910.

Qiu, B., Huang, Z., Liu, X., Meng, X., You, Y., Liu, G., Yang, K., Maier, A., Ren, Q., and Lu, Y. (2020). Noise reduction in optical coherence tomography images using a deep neural network with perceptually-sensitive loss function. *Biomedical optics express*, 11(2):817–830.

Qiu, B., You, Y., Huang, Z., Meng, X., Jiang, Z., Zhou, C., Liu, G., Yang, K., Ren, Q., and Lu, Y. (2021). N2nsr-oct: Simultaneous denoising and super-resolution in optical coherence tomography images using semisupervised deep learning. *Journal of Biophotonics*, 14(1):e202000282.

Rao, H. L., Pradhan, Z. S., Suh, M. H., Moghimi, S., Mansouri, K., and Weinreb, R. N. (2020). Optical coherence tomography angiography in glaucoma. *Journal of glaucoma*, 29(4):312–321.

Ren, M., Dey, N., Fishbaugh, J., and Gerig, G. (2021). Segmentation-renormalized deep feature modulation for unpaired image harmonization. *IEEE Transactions on Medical Imaging*, 40(6):1519–1530.

Reza, A. M. (2004). Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38:35–44.

Rico-Jimenez, J. J., Hu, D., Tang, E. M., Oguz, I., and Tao, Y. K. (2022). Real-time oct image denoising using a self-fusion neural network. *Biomedical Optics Express*, 13(3):1398–1409.

Rim, T. H., Teo, A. W. J., Yang, H. H. S., Cheung, C. Y., and Wong, T. Y. (2020). Retinal vascular signs and cerebrovascular diseases. *Journal of Neuro-ophthalmology*, 40(1):44–59.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Sagheer, S. V. M. and George, S. N. (2020). A review on medical image denoising algorithms. *Biomedical signal processing and control*, 61:102036.

Sakamoto, A., Hangai, M., and Yoshimura, N. (2008). Spectral-domain optical coherence tomography with multiple b-scan averaging for enhanced imaging of retinal diseases. *Ophthalmology*, 115(6):1071–1078.

Sander, B., Larsen, M., Thrane, L., Hougaard, J. L., and Jørgensen, T. M. (2005). Enhanced optical coherence tomography imaging by multiple scan averaging. *British Journal of Ophthalmology*, 89(2):207–212.

Sasongko, M., Wong, T., Nguyen, T., Cheung, C., Shaw, J., and Wang, J. (2011). Retinal vascular tortuosity in persons with diabetes and diabetic retinopathy. *Diabetologia*, 54:2409–2416.

Schmitt, J. (1997). Array detection for speckle reduction in optical coherence microscopy. *Physics in Medicine & Biology*, 42(7):1427.

Schmitt, J. M., Xiang, S., and Yung, K. M. (1999). Speckle in optical coherence tomography: an overview. In *Saratov Fall Meeting'98: Light Scattering Technologies for Mechanics, Biomedicine, and Material Science*, volume 3726, pages 450–461. International Society for Optics and Photonics.

Schreur, V., Domanian, A., Liefers, B., Venhuizen, F. G., Klevering, B. J., Hoyng, C. B., de Jong, E. K., and Theelen, T. (2019). Morphological and topographical appearance of microaneurysms on optical coherence tomography angiography. *British Journal of Ophthalmology*, 103(5):630–635.

Selim, M., Zhang, J., Fei, B., Zhang, G.-Q., and Chen, J. (2021). Ct image harmonization for enhancing radiomics studies. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1057–1062. IEEE.

Shi, F., Cai, N., Gu, Y., Hu, D., Ma, Y., Chen, Y., and Chen, X. (2019). Despecnet: a cnn-based method for speckle reduction in retinal optical coherence tomography images. *Physics in Medicine & Biology*, 64(17):175010.

Shu, Y., Cao, Z., Wang, C., Wang, J., and Long, M. (2021). Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9624–9633.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Spaide, R. F., Klancnik, J. M., and Cooney, M. J. (2015). Retinal vascular layers imaged by fluorescein angiography and optical coherence tomography angiography. *JAMA ophthalmology*, 133(1):45–50.

Srinivasan, V. J., Wojtkowski, M., Witkin, A. J., Duker, J. S., Ko, T. H., Carvalho, M., Schuman, J. S., Kowalczyk, A., and Fujimoto, J. G. (2006). High-definition and 3-dimensional imaging of macular pathologies with high-speed ultrahigh-resolution optical coherence tomography. *Ophthalmology*, 113(11):2054–2065.

Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A., and Van Ginneken, B. (2004). Ridge-based vessel segmentation in color images of the retina. *IEEE TMI*.

Tang, F. Y., Ng, D. S., Lam, A., Luk, F., Wong, R., Chan, C., Mohamed, S., Fong, A., Lok, J., Tso, T., et al. (2017). Determinants of quantitative optical coherence tomography angiography metrics in patients with diabetes. *Scientific reports*, 7(1):2575.

Toto, L., Borrelli, E., Di Antonio, L., Carpineto, P., and Mastropasqua, R. (2016). Retinal vascular plexuses'changes in dry age-related macular degeneration, evaluated by means of optical coherence tomography angiography. *Retina*, 36(8):1566–1572.

Toulouie, S., Chang, S., Pan, J., Snyder, K., Yiu, G., et al. (2022). Relationship of retinal vessel caliber with age-related macular degeneration. *Journal of Ophthalmology*, 2022.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Venkatanath, N., Praneeth, D., Bh, M. C., Channappayya, S. S., and Medasani, S. S. (2015). Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*, pages 1–6. IEEE.

Verma, K., Deep, P., and Ramakrishnan, A. (2011). Detection and classification of diabetic retinopathy using retinal images. In *2011 Annual IEEE India Conference*, pages 1–6. IEEE.

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., and Bengio, Y. (2019). Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR.

Virgili, G., Menchini, F., Casazza, G., Hogg, R., Das, R. R., Wang, X., and Michelessi, M. (2015). Optical coherence tomography (oct) for detection of macular oedema in patients with diabetic retinopathy. *Cochrane Database of Systematic Reviews*, (1).

Volpi, R. and Murino, V. (2019). Addressing model vulnerability to distributional shifts over image transformation sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7980–7989.

Walsh, J. B. (1982). Hypertensive retinopathy: description, classification, and prognosis. *Ophthalmology*, 89(10):1127–1131.

Wang, C., Chen, X., Ning, H., and Li, S. (2024). Sam-octa: A fine-tuning strategy for applying foundation model octa image segmentation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1771–1775. IEEE.

Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C., and Yushkevich, P. A. (2012). Multi-atlas segmentation with joint label fusion. *IEEE PAMI*, 35(3):611–623.

Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., and Li, H. (2022). Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Wei, X., Liu, X., Yu, A., Fu, T., and Liu, D. (2018). Clustering-oriented multiple convolutional neural networks for optical coherence tomography image denoising. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5. IEEE.

Wong, A., Mishra, A., Bizheva, K., and Clausi, D. A. (2010). General bayesian estimation for speckle noise reduction in optical coherence tomography retinal imagery. *Optics express*, 18(8):8338–8352.

Wong, T. Y. and McIntosh, R. (2005). Hypertensive retinopathy signs as risk indicators of cardiovascular morbidity and mortality. *British medical bulletin*, 73(1):57–70.

Wu, D., Gong, K., Kim, K., Li, X., and Li, Q. (2019). Consensus neural network for medical imaging denoising with only noisy training samples. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 741–749. Springer.

Xu, M., Tang, C., Hao, F., Chen, M., and Lei, Z. (2020a). Texture preservation and speckle reduction in poor optical coherence tomography using the convolutional neural network. *Medical Image Analysis*, 64:101727.

Xu, Z., Liu, D., Yang, J., Raffel, C., and Niethammer, M. (2020b). Robust and generalizable visual representation learning via random convolutions. *arXiv preprint arXiv:2007.13003*.

Yadav, S. S. and Jadhav, S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big data*, 6(1):1–18.

Yang, H., Wang, Z., Liu, X., Li, C., Xin, J., and Wang, Z. (2023). Deep learning in medical image super resolution: a review. *Applied Intelligence*, 53(18):20891–20916.

Yannuzzi, L. A., Rohrer, K. T., Tindel, L. J., Sobel, R. S., Costanza, M. A., Shields, W., and Zang, E. (1986). Fluorescein angiography complication survey. *Ophthalmology*, 93(5):611–617.

Yao, X., Liu, H., Hu, D., Lu, D., Lou, A., Li, H., Deng, R., Arenas, G., Oguz, B., Schwartz, N., et al. (2024). Fnpc-sam: uncertainty-guided false negative/positive control for sam on noisy medical images. In *Medical Imaging 2024: Image Processing*, volume 12926, page 1292602. SPIE.

You, Q., Freeman, W. R., Weinreb, R. N., Zangwill, L., Manalastas, P. I., Saunders, L. J., and Nudleman, E. (2017). Reproducibility of vessel density measurement with optical coherence tomography angiography in eyes with and without retinopathy. *Retina*, 37(8):1475–1482.

Yu, A., Liu, X., Wei, X., Fu, T., and Liu, D. (2018). Generative adversarial networks with dense connection for optical coherence tomography images denoising. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5. IEEE.

Yu, X., Li, G., Lou, W., Liu, S., Wan, X., Chen, Y., and Li, H. (2023). Diffusion-based data augmentation for nuclei image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 592–602. Springer.

Yushkevich, P. A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J. C., and Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, J., Qiao, Y., Sarabi, M. S., Khansari, M. M., Gahm, J. K., Kashani, A. H., and Shi, Y. (2019). 3d shape modeling and analysis of retinal microvasculature in oct-angiography images. *IEEE transactions on medical imaging*, 39(5):1335–1346.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. (2020a). How does mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*.

Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B. J., Roth, H., Myronenko, A., Xu, D., et al. (2020b). Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Zuo, L., Dewey, B. E., Liu, Y., He, Y., Newsome, S. D., Mowry, E. M., Resnick, S. M., Prince, J. L., and Carass, A. (2021). Unsupervised mr harmonization by learning disentangled representations using information bottleneck theory. *NeuroImage*, 243:118569.