UNIVERSITY OF CALIFORNIA SAN DIEGO

Deep Learning Based Multimodal Retinal Image Processing

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

in

Electrical Engineering (Signal and Image Processing)

by

Yiqian Wang

Committee in charge:

Professor Truong Nguyen, Chair
Professor Cheolhong An, Co-Chair
Professor William Freeman
Professor Ravi Ramamoorthi
Professor Nuno Vasconcelos

2022

The Dissertation of Yiqian Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Professor Truong Nguyen and Professor Cheolhong An for their support as the chair and co-chair of my committee. I deeply appreciate their continuous support and guidance during my PhD study. I would also like to thank my committee members for their invaluable advice on my dissertation work. I would like to thank members of Video Processing Lab and Jacobs Retinal Center for their technical support on my study. Last but not the least, I wish to extend my special thanks to my husband, parents, family and friends, without whose support I couldn't have gone so far.

Chapter 2, in part, is a reprint of the material as it appears in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Y. Wang, J. Zhang, C. An, M. Cavichini, M. Jhingan, M. J. Amador-Patarroyo, C. P. Long, D. G. Bartsch, W. R. Freeman and T. Q. Nguyen, IEEE, 2021, and in IEEE Transactions on Image Processing, Y. Wang, J. Zhang, M. Cavichini, D. G. Bartsch, W. R. Freeman, T. Q. Nguyen, C. An, IEEE, 2021. The dissertation author is the primary author of the two papers.

Chapter 3, in part, is a reprint of the material as it appears in IEEE Access, Y. Wang, J. Zhang, M. Cavichini, D. G. Bartsch, W. R. Freeman, T. Q. Nguyen, C. An. The dissertation author is the primary author of the this paper.

Chapter 4, in part, is a reprint of the material as it appears in IEEE International Conference on Image Processing 2021 (ICIP), Y. Wang, A. Warter, M. Cavichini, W. R. Freeman, D. G. Bartsch, T. Q. Nguyen, C. An, IEEE, 2021. The dissertation author is the primary author of this paper. Chapter 4, in part, has been submitted for publication of the material as it may appear in IEEE Transactions on Image Processing, Y. Wang, A. Warter, M. Cavichini, V. Alex, D. G. Bartsch, W. R. Freeman, T. Q. Nguyen, C. An, IEEE, 2021. The dissertation author is the primary author of this paper.

Chapter 5, in part, has been submitted for publication of the material as it may appear in IEEE International Conference on Image Processing 2022 (ICIP), Y. Wang, C. Galang, W. R. Freeman, T. Q. Nguyen, C. An, IEEE, 2022. The dissertation author is the primary author of this

paper. Chapter 5, in part, is currently being prepared for submission for publication of the material. Y. Wang, C. Galang, A. Warter, A. Heinke, D. G. Bartsch, W. R. Freeman, T. Q. Nguyen, C. An. The dissertation author is the primary author of this paper.

VITA

2014 - 2018 Bachelor of Science, in Electrical Engineering,
Beijing Institute of Technology

2018 - 2022 Doctor of Philosophy, in Electrical Engineering (Signal and Image Processing),
University of California San Diego

PUBLICATIONS

1. **Y. Wang**, C. Galang, W. R. Freeman, A. Warter, A. Heinke, D. -U. G. Bartsch, T. Q. Nguyen, and C. An, "Retinal OCT layer segmentation via joint motion correction and graph-assisted 3D neural network," in *IEEE Transactions on Image Processing*, (submitted).
2. **Y. Wang**, C. Galang, W. R. Freeman, T. Q. Nguyen, and C. An, "Joint motion correction and 3D segmentation with graph-assisted neural networks for retinal OCT," in *IEEE International Conference on Image Processing 2022 (ICIP)*, (submitted).
3. **Y. Wang**, A. Warter, M. Cavichini, V. Alex, D.-U. G. Bartsch, W. R. Freeman, T. Q. Nguyen, and C. An, "Deep learning network to correct axial and coronal eye motion in 3D OCT retinal imaging," *IEEE Transactions on Image Processing*, (submitted).
4. C. An, **Y. Wang**, J, Zhang, and T. Q. Nguyen, "Self-Supervised Rigid Registration for Multimodal Retinal Images," *IEEE Transactions on Image Processing*, (submitted).
5. J. Zhang, **Y. Wang**, J. Dai, M. Cavichini, D.-U. G. Bartsch, W. R. Freeman, T. Q. Nguyen, and C. An, "Two-Step Registration on Multi-Modal Retinal Images via Deep Neural Networks," *IEEE Transactions on Image Processing*, vol. 31, pp. 823-838, Dec. 2021.
6. J. Zhang, **Y. Wang**, D.-U. G. Bartsch, W. R. Freeman, T. Q. Nguyen, and C. An, "Perspective distortion correction for multi-modal registration between ultra-widefield and narrow-angle retinal images," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine*; Biology Society (EMBC), 2021, pp. 4086-4091.
7. C.J. Ho, **Y. Wang**, J. Zhang, T. Nguyen, and C. An, "A Convolutional Neural Network Pipeline For Multi-Temporal Retinal Image Registration", in *2021 18th International SoC Design Conference (ISOCC)*, 2021. pp. 27-28.
8. **Y. Wang**, A. Warter, M. Cavichini, W. R. Freeman, D. G. Bartsch, T. Q. Nguyen, C. An, "Learning to correct axial motion in OCT for 3D retinal imaging", in *IEEE International Conference on Image Processing 2021 (ICIP)*, 2021, pp. 126-130.
9. **Y. Wang**, J. Zhang, M. Cavichini, D. G. Bartsch, W. R. Freeman, T. Q. Nguyen, C. An, "Robust content-adaptive global registration for multimodal retinal images using weakly supervised deep-learning framework", *IEEE Transactions on Image Processing*, vol. 30, pp. 3167-3178, Feb. 2021.
10. **Y. Wang**, J. Zhang, M. Cavichini, D. G. Bartsch, W. R. Freeman, T. Q. Nguyen, C. An, "Study on correlation between subjective and objective metrics for multimodal retinal image registration", *IEEE Access*, vol. 8, pp. 190897-190905, Oct. 2020.

11. C. An, **Y. Wang**, J. Zhang, et al, "Fovea localization neural network for multimodal retinal imaging", *Applications of Machine Learning 2020*. International Society for Optics and Photonics, vol. 11511, pp. 196-202, Aug. 2020.
12. **Y. Wang**, J. Zhang, C. An, M. Cavichini, M. Jhingan, M. J. Amador, C. P. Long, D. G. Bartsch, W. R. Freeman, T. Q. Nguyen, "A segmentation based robust deep learning framework for multi-modal retinal image registration", in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 1369-1373.
13. **Y. Wang** and B. Li, "The fractional Fourier transform on graphs: sampling and recovery", in Proceedings of *2018 14th IEEE International Conference on Signal Processing*, Beijing, China, 2018, pp. 1103-1108.
14. **Y. Wang**, B. Li, Q. Cheng, "The fractional Fourier transform on graphs", in Proceedings of *Asia Pacific Signal and Information Processing Association Annual Summit and Conference 2017*, Kuala Lumpur, Malaysia, 2017, pp. 105-110.

ABSTRACT OF THE DISSERTATION

Deep Learning Based Multimodal Retinal Image Processing

by

Yiqian Wang

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

University of California San Diego, 2022

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

The retina, the light sensitive tissue lining the interior of the eye, is the only part of the central nervous system (CNS) that can be imaged at micron resolution in vivo. Retinal diseases including age-related macular degeneration, diabetes retinopathy, and vascular occlusions are important causes of vision loss and have systemic implications for millions of patients. Retinal imaging is of great significance to diagnosing and monitoring both retinal diseases and systematic diseases that manifest in the retina. A variety of imaging devices have been developed, including color fundus (CF) photography, infrared reflectance (IR), fundus autofluorescence (FAF), dye-based angiography, optical coherence tomography (OCT), and OCT angiography (OCT-A). Each imaging modality is particularly useful for observing certain aspects of the retina, and can be utilized for

visualization of specific diseases.

In this dissertation, we propose deep learning based methods for retinal image processing, including multimodal retinal image registration, OCT motion correction, and OCT retinal layer segmentation. We present our established work on a deep learning framework for multimodal retinal image registration, a comprehensive study of the correlation between subjective and objective evaluation metrics for multimodal retinal image registration, convolutional neural networks for correction of axial and coronal motion artifacts in 3D OCT volumes, and joint motion correction and 3D OCT layer segmentation network.

The dissertation not only proposes novel approaches in image processing, enhances the observation of retinal diseases, but will also provide insights on observing systematic diseases through the retina, including diabetes, cardiovascular disease, and preclinical Alzheimer's Disease. The proposed deep learning based retinal image processing approaches would build a connection between ophthalmology and image processing literature, and the findings may provide a good insight for researchers who investigate retinal image registration, retinal image segmentation and retinal disease detection.

Chapter 1

Introduction

The retina, the light sensitive tissue lining the interior of the eye, is the only part of the central nervous system (CNS) that can be imaged at micron resolution in vivo [1]. The retinal circulation is also part of the CNS circulation, and CNS diseases can be commonly observed in the retina. Therefore, the retina is truly the window to the brain, and retinal diseases are usually strongly connected to systemic disorders, such as emboli, strokes, diabetes, and so on.

Retinal imaging and the corresponding image processing techniques are of great significance to diagnosing and monitoring both retinal diseases and systematic diseases that manifest in the retina [2]. A variety of imaging devices have been developed, including color fundus (CF) photography, infrared reflectance (IR), fundus autofluorescence (FAF), dye-based angiography, optical coherence tomography (OCT), and OCT angiography (OCT-A). Each imaging modality is particularly useful for observing certain aspect of the retina, and can be utilized for visualization of specific diseases [3]. Therefore in clinical tests, multiple images are usually captured from multiple instruments to synthesize their complementary information.

Use of artificial intelligence (AI), and deep learning in particular, is gradually starting to grow in retinal analytics and does show promising results [4, 5, 6, 7, 8, 9]. In this work, we propose deep learning-based methods for retinal image processing, including multimodal retinal image registration, OCT motion correction, and OCT retinal layer segmentation. The proposed research will not only benefit the observation and diagnosis of retinal pathologies in clinical applications, but also build a connection between ophthalmology and image processing literature, providing insight

for researchers who investigate retinal image registration, retinal image segmentation and image domain transformation.

1.1 Background

1.1.1 The retina

The retina is a layered tissue that lines the back of the eye, playing a similar role to that of the film or image sensor in a camera. As shown in Fig. 1.1, The light is first passed through the cornea, the anterior chamber, the pupil, the lens, the vitreous, and then finally focused on the retina. The ocular structure of the eye forms a two-dimensional image of the outside world on the retina, and the latter converts the incoming light into neural signals that enables further processing in the visual cortex of the brain to form visual perception [2].

Illustrated in Fig. 1.1, the retina is composed of several layers of neurons connected to each other by synapses and supported by outer layers of pigment epithelium, choroid, and the sclera. Photoreceptor cells, including cones and rods, are the major light-sensing cells in the retina. Cones usually function in good light conditions providing color perception and high-acuity vision used for tasks including reading, whereas rods operate in low light conditions providing luminance information and peripheral vision. The third type of light-sensitive cells, photosensitive ganglion cells, are important for entrainment of circadian rhythm and reflexive responses such as pupillary light reflex.

The retinal region can be commonly divided into eleven layers from inner to outer retina [2]:

1. Internal limiting membrane (ILM): boundary between retina and vitreous;
2. Retinal nerve fiber layer (RNFL): axons of ganglion cells;
3. Ganglion cell layer (GCL): cell bodies of ganglion cells;
4. Inner plexiform layer (IPL): axons of bipolar cells;
5. Inner nuclear layer (INL): cell bodies of bipolar and horizontal cells;

6. Outer plexiform layer (OPL): dendrites of horizontal cells and inner segments of rods and cones;
7. Outer nuclear layer (ONL): cell bodies of the photoreceptor cells;
8. External limiting membrane (ELM): junctional complexes between photoreceptor and supportive Müller cells;
9. Photoreceptor layers (PR): photosensitive outer segments of rods and cones;
10. Retinal pigment epithelium (RPE): pigmented cell layer;
11. Bruch's membrane (BM): the innermost layer of the choroid;

1.1.2 Retinal manifestations of eye and systemic diseases

Many prevalent diseases originated in the eye, the CNS or the cardiovascular system, can manifest in the retina. The following diseases are a few important diseases that can be analyzed by retinal imaging.

Diabetic Retinopathy (DR) is a complication of diabetes that gradually damages the retina. DR is the leading cause of blindness for the working age in the U.S. [11]. High glucose in bloodstream damages the retinal vessels, which may cause ischemia, leading to new vessel growths, which may then bleed and/or cause retinal detachment; or breakdown of the blood-retinal barrier, resulting in fluid leakage, diabetic macular edema (DME), and damage to photoreceptor cells [2].

Age-Related Macular Degeneration (AMD) is the most common cause of blindness in the U.S. for age group over 50, characterized by blurred or no vision in the center of the visual field [12]. AMD includes both dry and wet form, where the dry AMD that makes up 70-90% of AMD. In dry AMD, small white or yellow deposits, called drusen, form beneath the macula and causes progressive loss of visual acuity. The wet form, also known as choroidal neovascularization (CNV), leads to rapid irreversible vision loss. Wet AMD involves growth of new choroidal vessels through a break in the BM into the RPE, causing abnormal fuid collection below the retina.

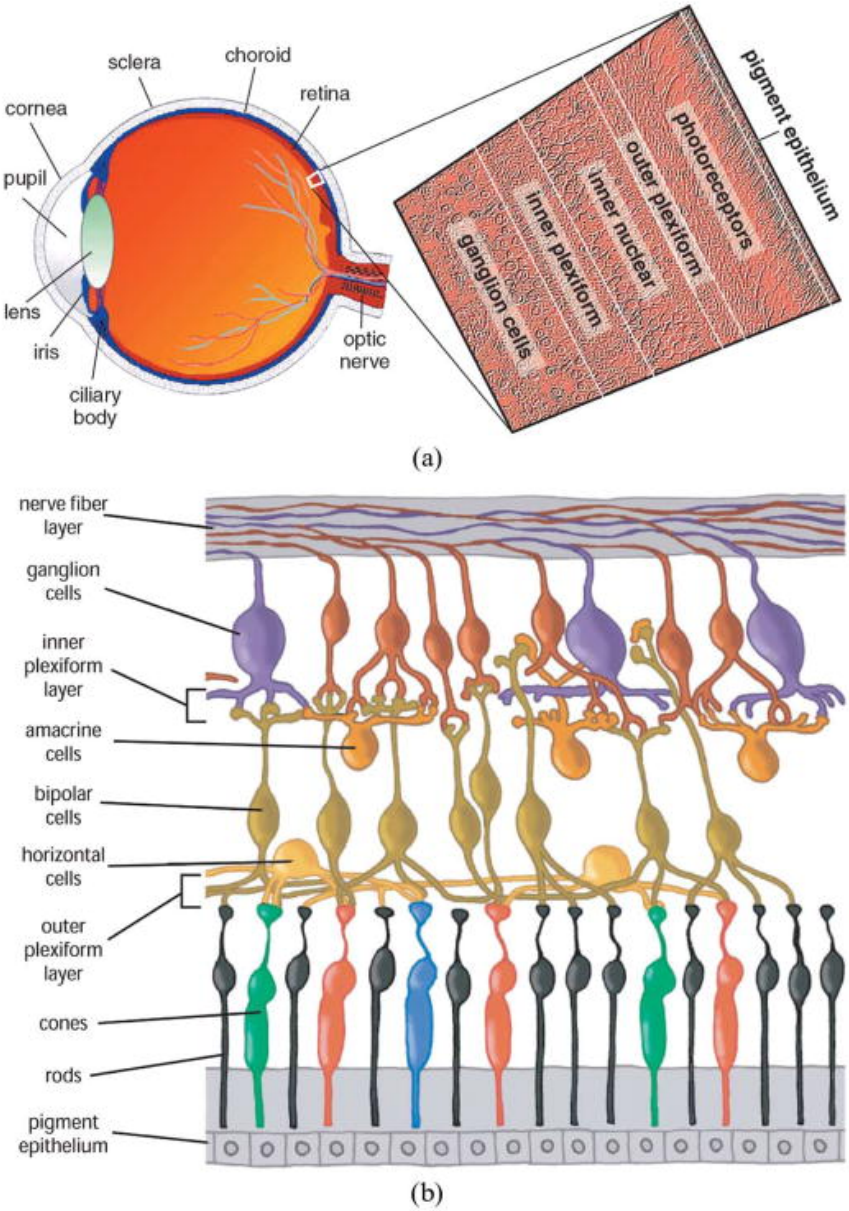


Figure 1.1. Eye anatomy and retinal layers [10]. (a) Diagram of a human eye, (b) cells in the retina arrayed in layers.

Glaucoma is the third most common cause of blindness in the U.S. [2], that gradually damages the optic nerve and loss of vision. Glaucoma is related to death of ganglion cells, and can be observed by cupping of the optic disc in retinal imaging. There are two categories including open-angle and closed-angle glaucoma. The open-angle glaucoma is known as the "silent thief of sight" that progresses slowly over time [13], while the closed-angle can be sudden and painful.

Cardiovascular Disease can manifest in the retina in various ways. For example, hypertension and atherosclerosis cause changes in A/V ratio in the retina, that is, the ratio between diameters of arteries and veins. An increased risk of myocardial infarction can be indicated by a decreasing A/V ratio [14]. Besides that, systemic vascular diseases can also cause central and branch artery and venous occlusions (CRAO, BRAO, CRVO, and BRVO).

Central nervous system (CNS) disorders can also manifest in the retina. The retina and optic nerve are considered part of the CNS, not only because they are extended from the diencephalon during embryonic development, but also due to the fact that fibers of the retinal ganglion cells, whose axons form the optic nerve, are actually CNS axons [1]. In addition, many retinal pathologies share similar features of other CNS pathologies, and CNS disorders can be detected and characterized through the retina in ophthalmological assessments, such as stroke, multiple sclerosis (MS), Parkinson disease (PD) and Alzheimer disease (AD) [1, 15].

1.1.3 Multimodal retinal imaging

As the optics of the eye is transparent to be imaged from the outside world, it is possible to observe the retina non-invasively from the outside with proper imaging technique. A variety of imaging devices have been developed, including color fundus (CF) photography, infrared reflectance (IR), dye-based angiography, optical coherence tomography (OCT), and OCT angiography (OCT-A).

The retinal imaging technologies can be categorized as *structural tests* and *functional tests*. Structural tests reveal the anatomy of retinal tissues, whereas functional tests can be performed to evaluate functionalities of tissues, which can be very helpful for detecting fluid leakage or vascular

occlusions. There are several important structural tests, including:

Color Fundus (CF) images are the gold standard for clinical documentation in most retinal diseases. A dedicated low-power microscope with attached white flash is used for capturing CF images with field of view ranging from 25 to 60-degree, visualizing central and peripheral retina, optic disc, and macula. Normal retina appears orange in CF images due to complexes of vitamin A with opsin proteins in the retina. Importantly, because white floodlights are used to obtain color fundus images, they are more representative of the clinician's visual impression than other imaging methods. Specially, true color of retina is useful to document vascular changes such as retinal hemorrhages, in diabetic retinopathy [16], hypertension, HIV disease [17] and other major medical problems.

Infrared Reflectance (IR) images, sometimes also referred to as infrared scaning laser ophthalmoscope (SLO) images, are typically obtained using an infrared light source with confocal SLO. The reflectance intensity is measured at each point to obtain a grayscale en-face image of the fundus [18]. Non-invasive IR imaging does not cause discomfort to the patient, and can sometimes be performed without dilation. It should be noted that many other SLO imaging modalities are intrinsically co-localized with IR, including blue- or green-light Fundus Autofluorescence (FAF), multicolor reflectance, and Optical Coherence Tomography (OCT), since they can be captured by the same instrument.

Fundus Autofluorescence (FAF) imaging uses blue or green light to illuminate the retina, and certain cellular components will "glow" without fluorescent dye injection [19, 20]. The autofluorescence from the retina is captured in a grayscale image, whose characteristic patterns of autofluorescence are useful in detecting and monitoring AMD, retinitis pigmentosa, central serous chorioretinopathy (CSC), and macular dystrophies.

Optical Coherence Tomography (OCT) is a powerful technique for non-invasive 3D imaging of the retina at micrometer resolution, that has become the standard of care for assessing most retinal conditions [2]. In OCT imaging, the sample is probed with a low-coherent infrared beam and the depth of backscattered light along the beam axis is measured by interference. Sequential

cross-sectional images are acquired by raster-scanning through the sample, and a 3D volume can be formed by stacking the cross-sectional images. An IR image is usually taken concurrently with OCT for en-face location reference [21]. OCT imaging identifies retinal layers and measures retinal thickness, which is helpful to detect drusen in AMD [2], DME in diabetic retinopathy [22], glaucoma [13] and so on.

Common functional tests include the followings:

Fluorescein Angiography (FA) is a type of dye-based angiography and an invasive test that requires intravenous administration of Sodium Fluorescein dye and imaging procedure of 10–30 minutes [23]. The fluoresce in the blood flow reveals abnormal blood vessels and leakage of liquid in a grayscale image. FA is the gold standard for the detection of CNV as well as retinal neovascularization, and is often used to help distinguish retinal diseases and determine whether laser treatment is needed on the retina [24].

Indocyanine Green Angiography (ICGA) is also an invasive test to obtain angiography of the choroid. Although the imaging procedure is similar to FA, the indocyanine green dye of ICGA fluoresces in infrared light and enables observation of deeper layers in the choroid. ICGA is widely used for detecting CNV in wet AMD [25] , focal delays, and hyperpermeability in central serous chorioretinopathy (CSCR) [26].

OCT Angiography (OCT-A) is a novel non-invasive imaging technique to acquire 3D angiography of the microvasculature of the retina and the choroid. After measuring temporal intensity variations to consecutive OCT B-scans, a blood flow map is constructed [23]. The motion caused by eye or head movement between or within sequential OCT scans must be eliminated, otherwise the intensity difference between repeated B-scans may be misinterpreted as flow of blood cells.

1.2 Retinal image processing

Retinal images can be processed and analyzed by image processing techniques to help ophthalmologists in diagnosing and monitoring of retinal pathology. Some important applications

in fundus photography include retinal image registration [27, 28, 29, 30, 31], montage of image sequences [32, 33, 34], retinal vessel segmentation [35, 6, 7, 8, 31], retinal disease detection [36, 4], and so on. However, most analysis of retinal images evaluate only one type of imaging modality. Multimodal or multi-instrumental analysis with deep learning has not been addressed. The ability to use deep learning to align and overlay different types of retinal images is a major innovation.

Besides 2D fundus photography, visualization of real 3D imaging in OCT and OCT-A revolutionized retinal imaging. Developing advanced processing techniques is crucial to extract clinically relevant information from the ever-increasing volume of data. Existing works include segmentation of retinal layers [4], 3D segmentation of retinal vessels [37, 38, 39], registration of 3D volumes [40], and prediction of disease progression [5]. Since the noise and motion artifacts in OCT scans significantly limit the performance of existing image processing techniques, it is very important to develop algorithms to correct artifacts and to utilize 3D contextual information.

1.2.1 Multimodal registration

The diagnostic benefit of overlaying different imaging modalities in retinal diseases has recently been demonstrated by different groups [15, 41, 42, 43]. However, it was extremely time-consuming work to process the large data, because all of these groups used human specialists to manually align multimodal images. Therefore, image registration, which overlays accurately and automatically the retinal images across different modalities, is critical to help ophthalmologists interpret each specialized test.

This can be illustrated with the following example. Fig. 1.2 shows that a color fundus photo is aligned and overlaid onto an infrared image. From sub-Figures 1.2, the color image in (a) reveals red (blue arrow) and white lesions (yellow arrow) in a diabetic patient. Note that color fundus photography is a standard screening tool for population studies of diabetes. The infrared image in (b) is taken simultaneously with the OCT raster images in (e1) and (f1). As infrared SLO images are intrinsically aligned with other SLO modalities including OCT, overlaying an IR image onto the CF image enhances analysis of any specified lesion in single modality with all these SLO modalities.

It also builds an imaging matrix where each point of the retina has information from co-localized multiple instruments. The blue arrows point to the red lesion seen on the color fundus photograph (and by clinical examination), and the ability to co-localize it with infrared image (b) allows perfect scanning through the lesion using OCT scan. In this case, the red lesion is not a retinal hemorrhage (more common in hypertension) but rather a microaneurysm (more common in diabetes). The white lesion (yellow arrow) on the fundus photograph is not well seen on IR images but is confirmed as a cotton wool spot (superficial retinal infarction) by the second OCT in (f) (yellow arrow).

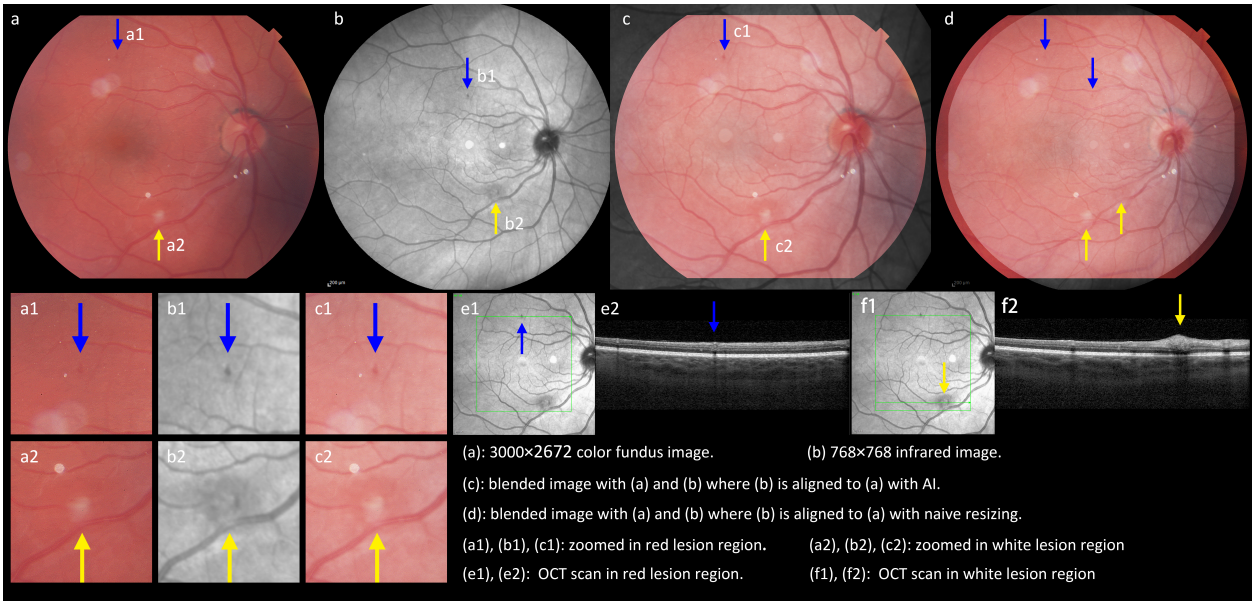


Figure 1.2. Alignment of color fundus (a), infrared image (b), and OCT scans (e), (f). The images reveal red lesion (blue arrow) and white lesion (yellow arrow) in a diabetic patient.

It is therefore important to design an algorithm to automatically align the retinal images across different modalities. However, it is complicated by the fact that the fields of view, lens systems, light sources and manufactures of such devices are all different. Fortunately, since the retinal vessels can be observed by all of these instruments in different ways, they are key features to align and overlay various diagnostic modalities.

Conventional approaches for multimodal retinal image registration can be roughly divided into two categories: area-based methods and feature-based methods. The *area-based methods* are designed to minimize the mutual information [44] or the entropy correlation coefficient [45]

between the images to be aligned, but they are sensitive to texture variations across modalities and also require intensive computation. The *feature-based methods* overcame this limitation by detecting sparse feature correspondences in both images to estimates the transformation. However, these conventional methods are still not robust enough for image pairs affected by disease or poor imaging quality, which covers up useful features for registration.

1.2.2 Eye motion correction

A major challenge in OCT imaging is the axial and horizontal motion artifacts introduced by involuntary eye movements. Even when the patient fixates upon a fixed object, the eye still carries out small and rapid movements including rapid microsaccades, high frequency tremors, and slow drifts with various frequency and magnitude [46]. Although high-speed Spectral Domain-OCT (SD-OCT) devices can acquire more than 300,000 A-scans per second [47], it typically takes several seconds to scan a 3D volume [48]. Therefore, artifacts due to the fixational eye movements are inevitable [46]. These involuntary eye movements would introduce both axial and coronal distortion, which leads to discontinuity of the 3D OCT data as shown in Fig. 1.3, where the motion artifacts are pointed by red arrows. It has also been shown that eye motion artifacts compromise subsequent analysis including retinal layer segmentation, OCT-A imaging, and detection of retinal diseases [49].

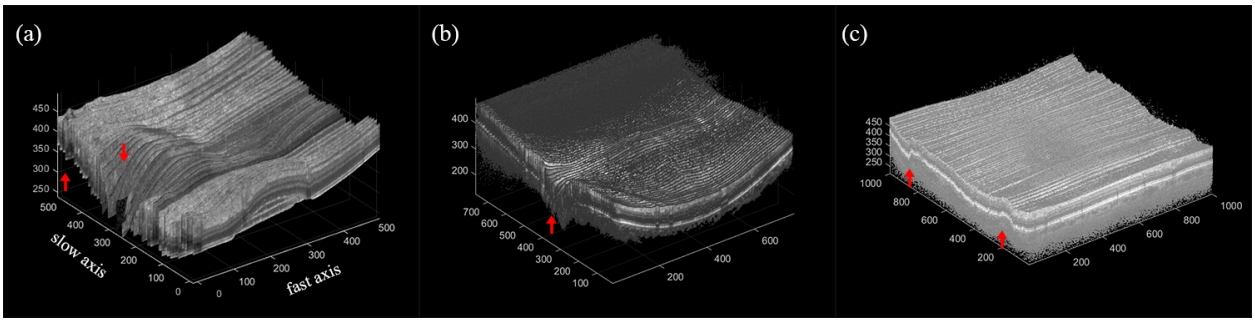


Figure 1.3. 3D OCT volume with motion artifacts. (a) Our dataset, (b) Dataset by Chiu et al. [50], (c) Dataset by Farsiu et al. [36].

Existing literature on OCT motion correction can be categorized into *prospective* and *retrospective* approaches [46]. Prospective approaches include hardware eye-tracking systems that

detect and compensate for motions during image acquisition [51, 52, 21]. Specifically, Heidelberg Spectralis [21] implemented the TruTrack Active Eye Tracking system, that tracks eye motion with two beams of light simultaneously. Although hardware-based methods can usually achieve more accurate results, they are difficult to implement and not available for every OCT device. Hence, software based retrospective approaches remains an active field of study.

Retrospective approaches are applied after image acquisition is done, and most are software-based methods. Most successful software methods require more than one OCT volumes [47, 53] or multimodal images as reference, that introduce extra burden for clinical tests. Other methods based on a single OCT volume tend to remove the curvature of the retina generating overly smoothed result [54, 55]. It is therefore important to develop an accurate motion correction algorithm that requires a single OCT volume and recovers the curvature of the retina.

1.2.3 Retinal layer segmentation

The retina is structured with multiple layers, and OCT segmentation aims to find the boundary between different layers automatically. As thickness changes of layers can indicate the status of diseases, layer segmentation will help ophthalmologists describe and locate lesions.

Most segmentation methods are categorized into 2D approaches, which is based on individual B-scans [56, 57, 58, 59, 50], and existing commercial OCT systems such as Heidelberg Spectralis [21] also integrate layer segmentation boundaries and thickness measurement features. Recently, deep learning methods significantly improved the accuracy of layer segmentation [60, 61, 4, 62, 63, 64, 65] using convolutional neural networks, especially U-net [66] which is the most popular structure for biomedical image segmentation in many recent works.

However, as shown in Fig. 1.4, most segmentation methods only operate on individual OCT B-scans to produce a segmentation map for each image slice, such that they ignore any 3D information contained in neighboring B-scans, which would be useful for segmentation [67]. This would yield discontinuity between B-scans when visualizing the layer segmentation surface in 3D. Instead of operating on single 2D OCT B-scans as most segmentation algorithms do, a neural

network can be trained directly on 3D OCT volume after motion correction to utilize contextual information and obtain more accurate layer segmentation labels. The 3D input will not only benefit the recognition of different tissue layers, but will also produce consistent segmentation boundaries throughout individual B-scans.

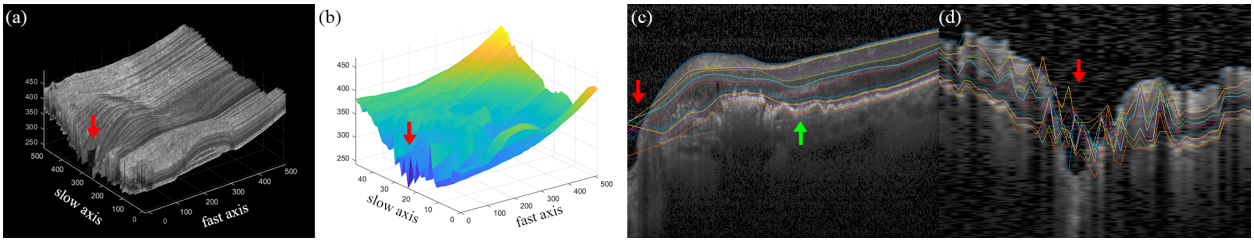


Figure 1.4. 2D OCT segmentation failures. (a) 3D OCT volume, (b) one example segmentation surface, (c) fast axis B-scan with layer segmentations, (d) cross-sectional slow axis B-scan with layer segmentations, red and green arrows denoting segmentation failures.

1.3 Contribution

The contributions of the dissertation are:

- A deep learning framework that focuses on the globally coarse alignment for the multimodal retinal image registration: It consists of three weakly supervised neural networks for vessel segmentation, feature detection and description, and outlier rejection. Experimental results demonstrate that the proposed framework can achieve state-of-the-art performance on two different datasets: public dataset and clinical dataset with a wide field of view. Specially, the framework is significantly robust to bad imaging quality compared to other methods.

- A comprehensive study on the existing evaluation metrics for multimodal retinal image registration, and a comparison of the Pearson correlation coefficient between the ophthalmologists' subjective grade and several commonly used objective evaluation metrics. Experimental results show that the Dice coefficient with deep learning-based segmentation is highly correlated with the subjective evaluation of the ophthalmologists.

- Convolutional neural networks that correct axial and coronal motion in OCT based on a single volume, while recovering the overall curvature of the retina. To the best of our knowledge, this

is the first fully deep learning-based method applied to the OCT motion correction problem, and experiment results show significant improvement compared with conventional methods in various resolutions an diseases.

- Combined motion correction and 3D neural network for retinal layer segmentation that utilizes 3D contextual information. The experimental result on three datasets shows the proposed method can achieve improved segmentation result compared with both conventional and deep learning methods, especially on a clinical dataset with severe diseases.

Chapter 2

Multimodal retinal image registration

2.1 Introduction

In the multimodal retinal image registration task, images captured from multiple imaging devices and modalities are aligned pixel-to-pixel to build a comprehensive co-localized database of the eye, and help ophthalmologists confirming their diagnosis with multiple evidences. However, it is challenging to detect and match common patterns across modalities [30], because multimodal images not only significantly differ in appearance, but also be captured in different resolution and field of view. Besides, poor imaging quality is ubiquitous in clinical applications, which could also distort useful patterns for matching and increase the difficulty of registration. Therefore, it has been a widely studied topic to design a robust registration approach for multimodal retinal images.

Many methods follow a *coarse-to-fine* registration pipeline [30, 68, 69]. The retina itself is close to spherical shape, but the imaging area can be approximated by a plane and registered with rigid, affine, or perspective transformation in the *globally coarse alignment* step. The remaining error due to planar approximation can be corrected by a deformable registration method in the *locally fine alignment* step. The coarse alignment step is crucial for successful registration, because most fine alignment methods cannot correct large errors in the previous coarse alignment step. Therefore, in this research, we focus on the coarse alignment step using perspective transformation.

Most existing coarse alignment approaches fall into two categories. The first category is *area-based*, which aims to minimize the mutual information [44] or the entropy correlation coefficient

[45] between the two images. Since the area-based methods are computationally intensive, and the performance degrades when substantial texture difference exists, it is not suitable for multimodal registration.

The second category is *feature-based*, that is based on detecting feature points and finding point correspondences in both images. A feature-based pipeline often includes vessel extraction, feature detection and description, initial matching and outlier rejection [30, 27, 28, 29, 70, 71]. Generally, the conventional algorithms for each step may not be optimal and adaptive to different images, and the resulting registration pipeline is often not robust enough to poor imaging quality.

In this chapter, a content-adaptive registration framework is proposed for multimodal retinal images, which focuses on the global alignment stage and includes three weakly supervised neural networks for vessel segmentation, feature detection and description, and outlier rejection. Compared with [72], we propose a novel content-adaptive segmentation network and fine-tune the feature detection and description network that improves performance. We also include more thorough comparison with literature and show robustness of the proposed registration with an extensive analysis on the influence of image quality. We demonstrate that our proposed framework can achieve state-of-the-art performance on two different datasets. Specifically, the framework is significantly robust to bad imaging quality compared to other methods in the clinical dataset with a wide field of view.

2.2 Related works

A number of conventional multimodal retinal image registration methods follow the pipeline including vessel extraction (if any), interest point detection and feature description, initial matching and outlier rejection. Various frameworks can be designed by choosing different combinations with emphasis on different aspects and target applications.

2.2.1 Vessel segmentation

Many retinal image registration approaches first extract vascular information from the source and target images to enhance edges and corners in the vessel and unify the modality. For example, [71] used an edge map based on strip fitting, [30] extracted a mean phase image using the Reize transform and log-Gabor filters, and [73] explicitly generated the vessel segmentation to achieve more robust matching result.

In deep learning literature, DRIU [6], DUNet [7], and IterNet [8] obtained accurate vessel segmentation maps with strong supervision. That is, they are supervised by pixel-wise segmentation ground truth, which requires intensive manual annotation. Since there is no known dataset with ground truth vessel segmentation for IR retinal images, it is very challenging to directly apply them to our dataset with limited time and cost.

2.2.2 Feature detection and description

The commonly used feature detection and description methods in computer vision can also be utilized in the retinal registration task, including Harris corner detection [74], histogram of oriented gradients (HOG) [75], scale invariant feature transformation (SIFT) [76], and speeded up robust features (SURF) [77]. Particularly for the retinal image descriptor, [28] proposed a partial intensity invariant feature descriptor (PIIFD) that is suitable for multimodality, and [71] proposed a low-dimensional step pattern analysis algorithm (LoSPA) to improve the robustness for image pairs affected by disease.

DeepSPA [9] proposed a learning-based descriptor for retinal images, but it was trained based on the detected classes of hand-crafted step patterns using the conventional LoSPA descriptor [71], which could limit the performance of the network. The learning-based approaches for general images, including learned invariant feature transform (LIFT) [78] and universal correspondence network (UCN) [79], improved matching performance compared to the conventional methods [78, 79]. However, both methods are not very accurate, since descriptor and keypoints are not derived jointly. LIFT can generate descriptors after detecting keypoints, and UCN generated only

dense descriptors with each pixel considered as a keypoint. The SuperPoint network [80] overcame the limitation with one encoder and two decoders to obtain keypoints and descriptors jointly in a single forward pass, which outperformed LIFT and UCN [80]. However, the original model was trained first on synthetic dataset and then refined with natural images, which is not optimal for retinal images.

2.2.3 Outlier rejection

The most commonly used method for outlier detection in computer vision is Random Sample Consensus (RANSAC) [81], which is an iterative approach that randomly selects matching points and votes for models based on the number of inliers. Another popular method, least median of squares (LMEDS) [82] computes the median of square error in each iteration, but it is only robust when the inlier ratio is more than 50%. Other iterative methods such as PROSAC [83], R-RANSAC [84], and USAC [85] achieved only marginal improvement over RANSAC with more complex algorithms.

For the multimodal retinal image registration, iterative approaches such as GDB-ICP [27] and ED-DB-ICP [86] applied the refining bootstrap regions. However, they are sensitive to scale. Other methods include adaptive outlier rejection based on asymmetric Gaussian mixture model (AGMM) [70], or root mean square error with feature distance (RMSEFD) [87] , but the methods required longer runtime and intense tuning.

In deep learning literature, DSAC [88] introduced a differentiable RANSAC for end-to-end training with marginal improvement. By contrast, [89] trained a network to learn the prior knowledge of the problem and reject outlier matchings such that the network outperformed RANSAC by a significant margin. Since the network was designed to estimate the essential matrix for camera pose estimation, some modifications are necessary to estimate the perspective transformation matrix for image registration.

2.2.4 Learning-based image registration

Deep learning has been extensively used in the single-modal image registration task (more commonly known as *optical flow* estimation). However, most methods such as FlowNet [90, 91], PWC-Net [92], and the latest IRR-PWC [93] led to deformable registration for natural images. As several large scale synthetic image datasets are publicly available, these models can be extended to image pairs even with large displacements. For medical image registration, Voxelmorph [94], DIR-Net [95] and DLIR [69] were proposed to register single-modal images like computed tomography (CT) images, but their features are essentially different from those of retinal images, such that directly applying these methods may not have desirable result on retinal dataset. Furthermore, these methods are not suitable for multimodal registration.

Networks trained for multimodal retinal images [31, 96, 97] mostly focus on the deformable (local) registration step. They assume that the input pair have been affinely aligned or the field of view of multimodal retinal images are quite similar. If the deformable methods were directly applied to the original input images where large displacement exists or large difference of the field of view (45° color fundus, 30° IR as inputs in Fig 2.1), they would not be able to correctly align the images, which will be discussed in section 2.4. CNNGeo [68] is an end-to-end network for semantic alignment of multimodal natural images. It also follows the coarse-to-fine procedure by first estimating 6 parameters for affine transformation, and then 12 parameters for spline transformation. It could handle large displacements, but the success criterion in [68] is not as strict as that in retinal image registration [29, 71, 70].

2.3 Deep learning-based multimodal registration framework

In Fig. 2.1, we propose a framework for multimodal retinal image registration, that consists of a content-adaptive vessel segmentation network, a SuperPoint network, and an outlier rejection network. The vessel segmentation maps of the source and target images are first extracted by two separate content-adaptive vessel segmentation networks. Then the SuperPoint network [80] detects and describes features in both segmentation maps. The keypoints are matched using the

mutual nearest neighbor matching algorithm, and inlier matches are selected by the outlier rejection network. Finally, the perspective transformation matrix is calculated based on the inlier matches, and the source image is warped towards the target image.

The proposed method follows the feature based registration pipeline, and each step of the pipeline is optimized with minimal supervision (weakly supervised learning). To the best of our knowledge, the proposed method is the first fully learning-based approach for the global (rigid) registration of multimodal retinal images.

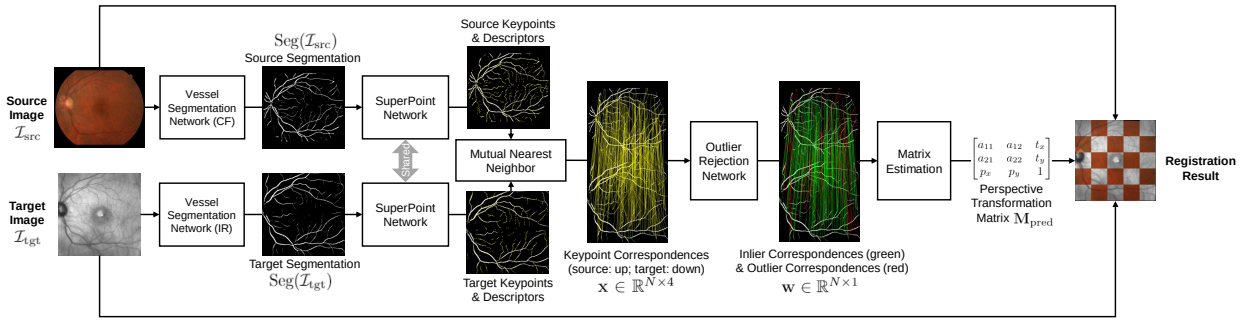


Figure 2.1. Block diagram of the proposed registration framework for multimodal retinal image registration, consisting of a content-adaptive vessel segmentation network, a SuperPoint network, and an outlier rejection network to estimate the perspective transformation matrix between source and target images.

2.3.1 Vessel segmentation network

The structure of the vessel segmentation network is shown in Fig. 2.2. The input image is RGB CF for source or IR for target. The single channel IR image is converted to three channels. The output segmentation map is a single-channel grayscale image with bright vessels and dark background. The source and target segmentation networks have separate encoders, but share the same decoder for segmentation output.

We propose the content-adaptive and unsupervised vessel segmentation network to overcome the dependency on manually labeled vessel segmentation ground truth, and to improve the robustness when aligning poor quality retinal images. To enable the network adaptive to different image contents, the pixel-adaptive convolution (PAC) [98] is adopted. Neural networks benefit from the weight-sharing nature of convolution whereas it is also content agnostic in the sense that the same

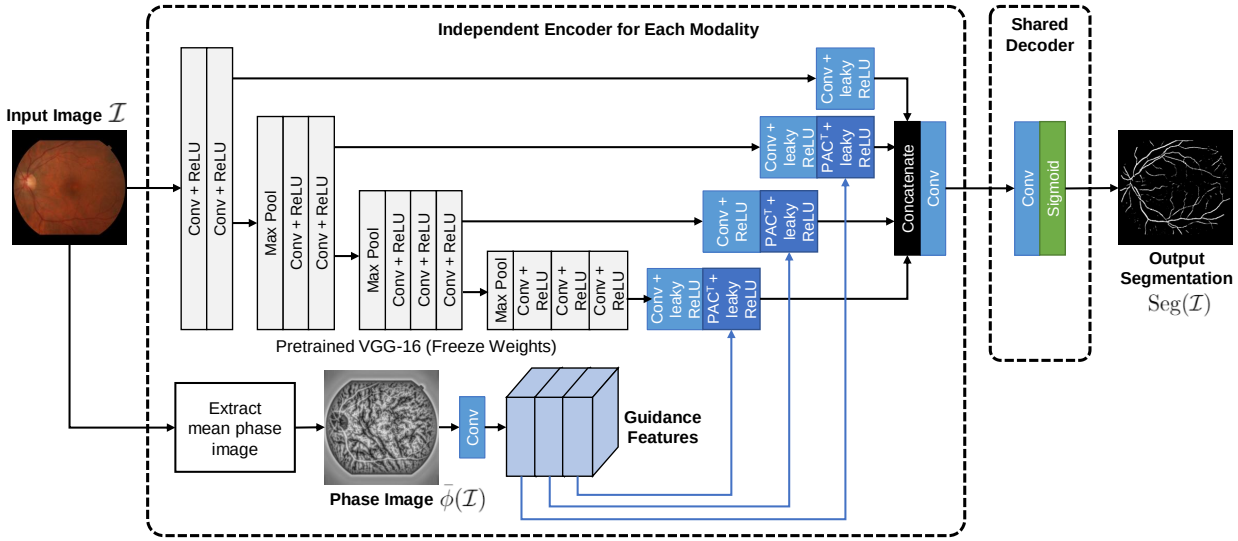


Figure 2.2. Structure of the content-adaptive vessel segmentation network. PAC^T stands for transposed pixel-adaptive convolution [98]

set of kernel is applied to different image contents and pixel locations. On the contrary, PAC [98] weighs the original 2D convolution kernel according to the feature guidance at different locations, such that the network can adapt to different content.

Similar to the example in [98], the upsampling transposed convolution blocks are replaced with transposed PAC (PAC^T), where we use the mean phase image [30] of the input as a guidance input, as illustrated in Fig. 2.2. The phase of image $\mathscr{I}$ at scale σ is first computed by

$$\phi_\sigma(\mathscr{I}) = \arctan\left(\frac{G_\sigma(|\mathbf{f_R}(\mathscr{I})|)}{G_\sigma(f_e(\mathscr{I}))}\right), \tag{2.1}$$

where $\mathbf{f_R}(\mathscr{I})$ is the odd component of $\mathscr{I}$ extracted by the Reize transform [99], $f_e(\mathscr{I})$ is the even component [100], and $G_\sigma()$ represents the log-Gabor filter at scale σ [101]. Then the mean phase image is obtained by taking the average of phase images at multiple scales,

$$\bar{\phi}(\mathscr{I}) = \frac{1}{N}\sum_{i=1}^{N}\phi_{\sigma_i}(\mathscr{I}). \tag{2.2}$$

Empirically, the mean phase image is very robust in enhancing edges in the original even with low contrast. The pixel adaptive convolution could adjust local convolution kernel weights in

the segmentation network according to the pixel intensities of guidance mean phase image, and thus help segmentation of vessels in bad quality retinal images. We will illustrate the benefits of content adaptation in detail in the experiment section 2.4.

To train the segmentation network in an unsupervised fashion, the style transfer technique is adopted [31]. Using style transfer, the network only requires one manually labeled segmentation map from any dataset to serve as a style reference. In this way, the network can be trained without pixel-wise segmentation ground truth for every image in the dataset, which would require extensive manual annotation. Especially, since expert annotated vessel segmentation of IR images is not publicly available, the unsupervised vessel segmentation network is a crucial step for the success of the proposed framework.

The loss function (2.7) is a hybrid loss, where the first term is designed to make the network mimic the style of the reference, while the last two terms regularize the network to keep the content of the input in the same position. The style loss [102] of the first term, $\mathscr{L}_{\text{style}}$ measures the style difference between the output segmentation and the style reference, which is a segmentation map labeled by hand from DRIVE dataset [103]. The global style of an image $\mathscr{I}$ is characterized by the output at different layers of a pretrained network ψ. Let $\psi_j(\mathscr{I})$ be the output feature at the j-th layer with shape $C_j \times H_j \times W_j$, the (i,k)-th element of the $C_j \times C_j$ Gram matrix is defined as

$$G_j(\mathscr{I})_{i,k} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \psi_j(\mathscr{I})_{i,h,w} \psi_j(\mathscr{I})_{k,h,w}. \tag{2.3}$$

Denoting the segmentation of image $\mathscr{I}$ with $\text{Seg}(\mathscr{I})$ and the style reference image $\mathscr{I}_{\text{ref}}$, then the style loss is defined as the squared Frobenius norm over their Gram matrices

$$\mathscr{L}_{\text{style}}(\mathscr{I}; \mathscr{I}_{\text{ref}}) = \sum_{j=1}^{N_L} ||G_j(\text{Seg}(\mathscr{I})) - G_j(\mathscr{I}_{\text{ref}})||_F^2. \tag{2.4}$$

The second term of (2.7) is a self comparison loss that compares the mean squared error (MSE) between the rotated segmentation of the rotated input and the segmentation of the original

input, denoted as

$$\mathscr{L}_{\text{self}}(\mathscr{I}) = \text{MSE}(\text{rot}(\text{Seg}(\text{rot}(\mathscr{I}))), \text{Seg}(\mathscr{I})), \tag{2.5}$$

where $\text{rot}(\mathscr{I})$ rotates the image $\mathscr{I}$ by 180°.

The last term is an image registration loss, which is a MSE on the aligned source and target segmentation using the manually labeled ground truth transformation matrix. Denote the source image by $\mathscr{I}_{\text{src}}$, the target image by $\mathscr{I}_{\text{tgt}}$, and the ground truth transformation matrix by $\mathbf{M}_{\text{GT}}$,

$$\mathscr{L}_{\text{reg}}(\mathscr{I}_{\text{src}}, \mathscr{I}_{\text{tgt}}; \mathbf{M}_{\text{GT}}) = \text{MSE}(\text{warp}(\text{Seg}(\mathscr{I}_{\text{src}}), \mathbf{M}_{\text{GT}}), \text{Seg}(\mathscr{I}_{\text{tgt}})). \tag{2.6}$$

The total loss is a weighted sum of the three terms

$$\begin{aligned} \mathscr{L}_{\text{seg}}(\mathscr{I}_{\text{src}}, \mathscr{I}_{\text{tgt}}; \mathscr{I}_{\text{ref}}, \mathbf{M}_{\text{GT}}) &= \lambda_{\text{style}}[\mathscr{L}_{\text{style}}(\mathscr{I}_{\text{src}}; \mathscr{I}_{\text{ref}}) + \mathscr{L}_{\text{style}}(\mathscr{I}_{\text{tgt}}; \mathscr{I}_{\text{ref}})] \\ &+ \lambda_{\text{self}}[\mathscr{L}_{\text{self}}(\mathscr{I}_{\text{src}}) + \mathscr{L}_{\text{self}}(\mathscr{I}_{\text{tgt}})] + \lambda_{\text{reg}}\mathscr{L}_{\text{reg}}(\mathscr{I}_{\text{src}}, \mathscr{I}_{\text{tgt}}; \mathbf{M}_{\text{GT}}), \end{aligned} \tag{2.7}$$

where λ_{style}, λ_{self}, and λ_{reg} are weighting parameters.

2.3.2 Feature detection and description network

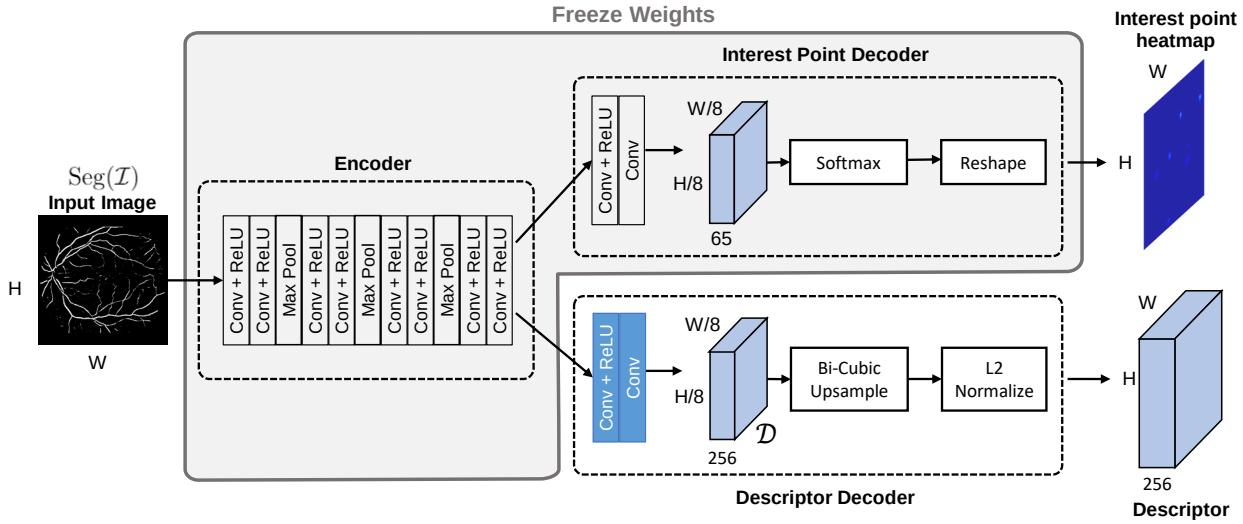


Figure 2.3. Structure of the fine-tuned Superpoint network at inference time.

The feature detection and description network is fine-tuned based on the pretrained Super-

Point model [80]. The structure of SuperPoint at inference time is shown in Fig. 2.3, which includes an encoder, an interest point decoder, and a descriptor decoder. The input is a single channel $H \times W$ image, and the output includes a $H \times W$ interest point heatmap and a $256 \times H \times W$ descriptor.

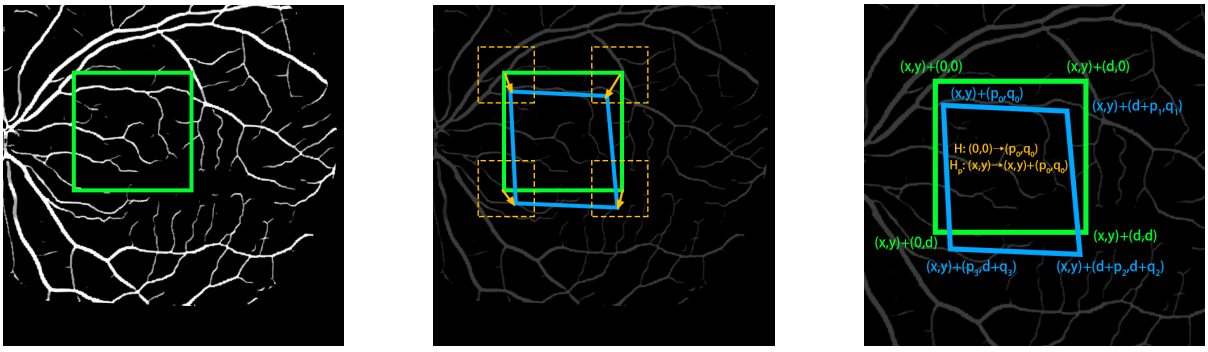
At training time, we freeze the weights in the encoder and the interest point decoder, and only fine-tune the descriptor decoder with the patches of the vessel segmentation map (input image) in our dataset. The procedure to crop patches and generate ground truth transformation matrix $\mathbf{H}$ is shown in Fig. 2.4. From sub-Figures 2.4, (a) the source segmentation image is warped using $\mathbf{M}_{\text{GT}}$, which is the same as the ground truth transformation matrix used in the previous section, and the source patch is cropped at a random position (x, y). (b) We then randomly perturb the four corners of the patch , and (c) compute the transformation matrices $\mathbf{H}$ and $\mathbf{H}_{\text{p}}$ from the corner coordinates. (d) we warp the target segmentation image using $\mathbf{H}_{\text{p}}$, and crop at the same position (x, y) to get the target patch. Finally, (e) we get the source and target patches, and the ground truth transformation matrix between the patches is $\mathbf{H}$.

The output of the descriptor decoder $\mathscr{D}$ in Fig. 2.3 is a semi-dense $256 \times \frac{H}{8} \times \frac{W}{8}$ tensor, where 256 dimensional descriptor is located at the center of each 8×8 cell in the original resolution. Let $\mathbf{p}_{ij} = [p_1, p_2]^{\text{T}}$ denote the center coordinate of the (i, j)th cell, and then the projected coordinate $T(\mathbf{p}_{ij}, \mathbf{H})$ of $\mathbf{p}_{ij}$ under the transformation $\mathbf{H}$ is

$$T(\mathbf{p}_{ij}, \mathbf{H}) = \begin{bmatrix} q_1/q_3 \\ q_2/q_3 \end{bmatrix}, \text{where} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} = \mathbf{H} \begin{bmatrix} p_1 \\ p_2 \\ 1 \end{bmatrix}. \tag{2.8}$$

We can obtain the ground truth label $s_{iji'j'}$ for correspondence between the (i, j)th cell in the source image patch and (i', j')th cell in the target image patch determined by the ground truth transformation matrix $\mathbf{H}$ between the two patches. Let $\mathbf{p}_{i'j'}$ be the center coordinate of the (i', j')th cell,

$$s_{iji'j'} = \begin{cases} 1, & \text{if } ||T(\mathbf{p}_{ij}, \mathbf{H}) - \mathbf{p}_{i'j'}|| \leq 5 \text{ pixels}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$



(a) Crop a source patch at random position (x,y) in the warped source image.

(b) Randomly perturb four corners of the patch.

(c) Compute transformation matrices $\mathbf{H}$ and $\mathbf{H}_p$.

(d) Warp the target image with $\mathbf{H}_p$. Crop a target patch at (x,y).

(e) The transformation matrix from the source patch to the target patch is $\mathbf{H}$.

Figure 2.4. Procedure to generate training patches for the SuperPoint network.

Here, we set the threshold to 5 pixels for better performance, instead of 8 pixels in [80]. The label for all possible correspondences between a total of $\frac{H}{8} \times \frac{W}{8}$ cells in the source image and $\frac{H}{8} \times \frac{W}{8}$ cells in the target image is denoted by S.

The descriptor loss $\mathscr{L}_{\text{desc}}$ in eq. (2.10) is a hinge loss to maximize the descriptor similarity between matching points ($s_{iji'j'} = 1$) and minimize the descriptor similarity between non-matching points ($s_{iji'j'} = 0$). The loss in eq. (2.11) is ignored when the correlation (dot product) of two matching descriptors is already larger than the positive margin m_p, or the correlation of two non-matching descriptors is already less than the negative margin m_n, in order to optimize for hard examples. The weighting factor λ_d is used to balance matching and non-matching examples. Let $\mathbf{d}_{i,j}$ denote the source descriptor at the (i,j)th cell and $\mathbf{d}'_{i',j'}$ denote the target descriptor at the (i',j')th

cell,

$$\mathscr{L}_{\text{desc}}(\mathscr{D}_{\text{src}}, \mathscr{D}_{\text{tgt}}; S) = \frac{1}{(HW/64)^2} \sum_{i=1}^{H/8} \sum_{j=1}^{W/8} \sum_{i'=1}^{H/8} \sum_{j'=1}^{W/8} l_d(\mathbf{d}_{ij}, \mathbf{d}'_{i'j'}; s_{iji'j'}), \tag{2.10}$$

$$l_d(\mathbf{d}, \mathbf{d}'; s) = \lambda_d s \max\{m_p - \mathbf{d}^T \mathbf{d}', 0\} + (1-s)\max\{\mathbf{d}^T \mathbf{d}' - m_n, 0\}. \tag{2.11}$$

At inference time, the interest point heatmap is post-processed by non-max suppression with the 5-pixel threshold, and keypoints are detected above the confidence threshold at 0.015. Then the corresponding descriptor vectors are matched by the mutual nearest neighbor algorithm, where the nearest neighbor from source to target must be the same as the nearest neighbor from target to source. After this step, we obtain the coordinates of keypoint correspondences $\mathbf{x}$ between source and target segmentation maps, as illustrated in Fig. 2.1.

2.3.3 Outlier rejection network

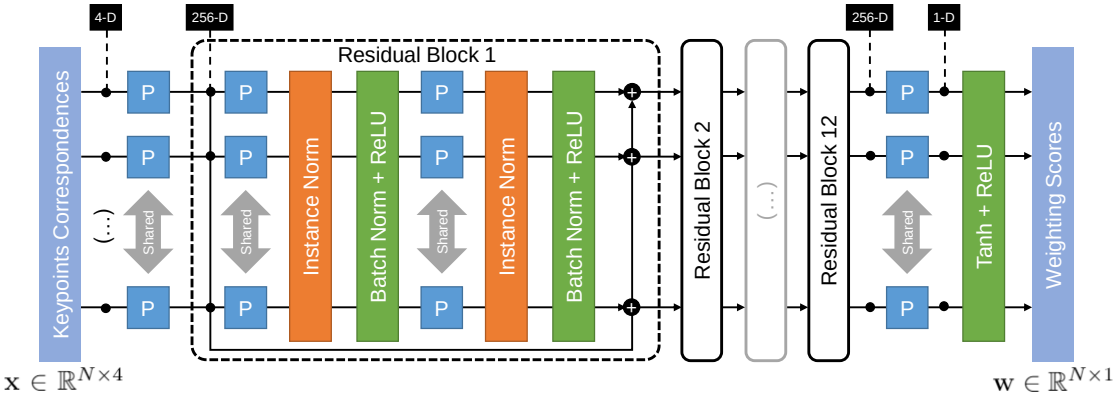


Figure 2.5. Structure of the outlier rejection network. "P" stands for Perceptron.

The outlier rejection network is shown in Fig. 2.5. The input $\mathbf{x} \in \mathbb{R}^{N\times 4}$ is N pairs of keypoint correspondences after the feature detection and description network, where $\mathbf{x}_i = [x_i, y_i, x'_i, y'_i]$, for $1 \leq i \leq N$. The output is a vector $\mathbf{w} \in \mathbb{R}^{N\times 1}$ containing a probability score $w_i \in [0,1)$ for each correspondence. The network consists of 12 residual blocks with 256 dimensional weight-sharing perceptrons and instance normalization, which is similar to the learned-correspondence network [89]. This structure guarantees that the output matrix is invariant to change of ordering in the input

point correspondences, as the ordering of the weighting scores also change correspondingly. The major difference to [89] is that we estimate the perspective transformation for image registration instead of the essential matrix used for camera pose estimation.

It requires at least 4 pairs of correspondences to estimate the perspective transformation that has 8 degrees of freedom. In order to estimate the 3×3 perspective transformation matrix $\mathbf{M}$ from pairs of coordinates and weighting scores, we adapt a weighted version of 4-point algorithm, which allows soft-assignment to put more weights on inliers with higher probabilities. The conventional 4-point algorithm uses hard-assignment, which is a special case of the weighted 4-point algorithm with binary weighting scores. For the i-th pair of correspondence, denote the source and the target coordinates by (x_i, y_i) and (x'_i, y'_i), respectively and define matrix $\mathbf{A} \in \mathbb{R}^{2N \times 9}$ as

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & x_1 & y_1 & 1 & -x_1 y'_1 & -y_1 y'_1 & -y'_1 \\ x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1 x'_1 & -y_1 x'_1 & -x'_1 \\ \vdots & & & & \vdots & & & & \vdots \\ 0 & 0 & 0 & x_N & y_N & 1 & -x_N y'_N & -y_N y'_N & -y'_N \\ x_N & y_N & 1 & 0 & 0 & 0 & -x_N x'_N & -y_N x'_N & -x'_N \end{bmatrix}, \tag{2.12}$$

then the problem is to find the vectorized transformation matrix $\mathrm{Vec}(\mathbf{M})$ that minimizes $||\mathbf{WA}\mathrm{Vec}(\mathbf{M})||$, where $\mathbf{W} \in \mathbb{R}^{2N \times 2N}$ is a diagonal matrix of the output scores $\mathbf{W} = \mathrm{diag}([w_1, w_1, \ldots, w_N, w_N])$. It can be proved that the solution is the corresponding eigenvector of the smallest eigenvalue of $\mathbf{A}^T \mathbf{W}^2 \mathbf{A}$.

For the outlier rejection network, the loss function $\mathscr{L}_{\text{outlier}}$ in eq. (2.19) is a weighted sum of three terms. The first term $\mathscr{L}_{\text{class}}$ is a classification loss, that is a cross-entropy loss between the predicted and ground truth labels for each correspondence (2.13). Let $o_i(\mathbf{x})$ be the last linear layer output for the i-th correspondence, and $y_i(\mathbf{M}_{\text{GT}}) \in \{0, 1\}$ be its label as an inlier given the ground truth transformation matrix $\mathbf{M}_{\text{GT}}$, then the classification loss can be defined as

$$\mathscr{L}_{\text{class}}(\mathbf{x}; \mathbf{M}_{\text{GT}}) = \frac{1}{N} \sum_{i=1}^{N} \gamma_i H(y_i(\mathbf{M}_{\text{GT}}), \sigma(o_i(\mathbf{x}))) \tag{2.13}$$

where $H(\cdot)$ denotes the binary cross-entropy, $\sigma(\cdot)$ denotes the sigmoid function, and γ_i is a per-label balancing factor. The label $y_i(\mathbf{M}_{\text{GT}})$ for each correspondence is obtained by thresholding the L_2 distance between the target coordinates $\mathbf{p}'_i$ and warped source coordinates $T(\mathbf{p}_i, \mathbf{M}_{\text{GT}})$ (2.8)

$$y_i(\mathbf{M}_{\text{GT}}) = \begin{cases} 1, & \text{if } ||T(\mathbf{p}_i, \mathbf{M}_{\text{GT}}) - \mathbf{p}'_i|| \leq 5 \text{ pixels} \\ 0, & \text{otherwise} \end{cases} \tag{2.14}$$

The matrix regression loss $\mathscr{L}_{\text{matrix}}$ is MSE between the predicted and ground thruth transformation matrix

$$\mathscr{L}_{\text{matrix}}(\mathbf{x}; \mathbf{M}_{\text{GT}}) = \text{MSE}(\mathbf{M}_{\text{GT}} - \mathbf{M}_{\text{pred}}(\mathbf{x})) \tag{2.15}$$

where $\mathbf{M}_{\text{pred}}(\mathbf{x})$ denotes the predicted transformation matrix given $\mathbf{x}$.

The Dice coefficient is commonly used to evaluate registration accuracy and the Dice loss is defined as one minus the soft Dice coefficient on the aligned segmentation images. The Dice coefficient for binary segmentation is defined as

$$\text{Dice}(\mathscr{I}_1, \mathscr{I}_2) = \frac{2 \times \sum(\mathscr{I}_1 \odot \mathscr{I}_2)}{\sum \mathscr{I}_1 + \sum \mathscr{I}_2} \tag{2.16}$$

where $\odot$ denotes element-wise product. It is a value between [0,1], and higher Dice coefficient indicates more overlap. Since the binary Dice coefficient is not differentiable, the soft Dice coefficient [31] is applied for grayscale segmentation maps as follows:

$$\text{Dice}_s(\mathscr{I}_1, \mathscr{I}_2) = \frac{2 \times \sum \text{ele\_min}(\mathscr{I}_1, \mathscr{I}_2)}{\sum \mathscr{I}_1 + \sum \mathscr{I}_2} \tag{2.17}$$

where ele_min denotes the element-wise minimum. Its registration Dice loss can then be written as

$$\mathscr{L}_{\text{Dice}}(\mathbf{x}, \mathscr{I}_{\text{src}}, \mathscr{I}_{\text{tgt}}) = 1 - \text{Dice}_s(\text{warp}(\mathscr{I}_{\text{src}}, \mathbf{M}_{\text{pred}}(\mathbf{x})), \mathscr{I}_{\text{tgt}}), \tag{2.18}$$

where $\mathscr{I}_{\text{src}}$ is the warped source segmentation and $\mathscr{I}_{\text{tgt}}$ is the target segmentation. Then the total

loss is

$$\begin{aligned}\mathscr{L}_{\text{outlier}}(\mathbf{x}, \mathscr{I}_{\text{src}}, \mathscr{I}_{\text{tgt}}; \mathbf{M}_{\text{GT}}) &= \lambda_{\text{class}} \mathscr{L}_{\text{class}}(\mathbf{x}; \mathbf{M}_{\text{GT}}) \\ &+ \lambda_{\text{matrix}} \mathscr{L}_{\text{matrix}}(\mathbf{x}; \mathbf{M}_{\text{GT}}) + \lambda_{\text{Dice}} \mathscr{L}_{\text{Dice}}(\mathbf{x}, \mathscr{I}_{\text{src}}, \mathscr{I}_{\text{tgt}}).\end{aligned} \tag{2.19}$$

2.4 Experimental result

In this experiment, we compare the overall performance of our proposed framework to several existing methods for multimodal retinal image registration that estimates the affine or perspective transformation.

2.4.1 CF-IR dataset

Dataset

The first dataset collected by Jacobs Retina Center (JRC) at Shiley Eye Institute consists of color fundus images (RGB, 3000×2672) for source and infrared reflectance (IR) images (grayscale, 768×768 or 1536×1536) for target. The image pairs contain a variety of pathologies including diabetes, hemorrhages, and macular degeneration. It is partitioned into 530 pairs for the training set, 90 for the validation set, and 253 for the test set. The ground truth transformation matrices $\mathbf{M}_{\text{GT}}$ for the training and validation set are calculated by manually labeled correspondences. The quality of each image in the dataset is rated by ophthalmologists as good, usable, and bad as listed in Table 2.1.

Table 2.1. Number of good, usable, and bad quality images in JRC CF-IR dataset.

Dataset	Good	Usable	Bad	Total
Training CF	256	193	81	530
Training IR	327	168	35	530
Validation CF	47	30	13	90
Validation IR	58	23	9	90
Test CF	124	90	40	253
Test IR	181	59	14	253

Criteria

The robustness of registration is measured by the success rate. We determine successful registration when the maximum error (MAE) in equation (2.20) is less than or equal to 10 pixels [28, 29, 70, 71, 9] on 6 manually labeled correspondences $\mathscr{P}$, as illustrated in Fig. 2.6.

$$\text{MAE} = \max_{\mathbf{p} \in \mathscr{P}} \left|\left| \, T\left(T\left(\mathbf{p}, \mathbf{M}_{\text{GT}}^{-1}\right), \mathbf{M}_{\text{pred}}\right) - \mathbf{p} \, \right|\right| \tag{2.20}$$

The accuracy of registration (regardless of success or not) is measured by the Dice coefficient (2.16) on the binary segmentation maps of aligned images, where the binary segmentation maps are obtained using our PAC vessel segmentation network with 0.5 threshold. We multiply the two segmentation maps with a valid mask to compute the Dice coefficient in their overlapping region.

Implementation

All images are first padded to square shape and resized to 768×768 before applying different registration methods. It is noticed that directly downsampling the source images will increase the noise and adversely affect the segmentation. Therefore, the images are anti-aliased with Gaussian filter before bicubic downsampling. In our method, all pixel intensities are normalized between $[0, 1]$, and the target grayscale images are converted to 3 channels by stacking the input channel 3 times.

Training

Three networks in the proposed framework are trained sequentially in PyTorch using Adam optimizer, and choose the best model with the lowest loss on the validation set. The content-adaptive vessel segmentation network is trained on our dataset using learning rate 10^{-3}, and batch size 1 due to limited memory. We used $\lambda_{\text{style}} = 1$, $\lambda_{\text{self}} = 10^{-4}$, and $\lambda_{\text{reg}} = 10^{-3}$ for the loss function (2.7), which yields the best result in our experiments. We observed that when λ_{self} and λ_{reg} are too large, the network may output images completely black, but when they are too small, the network would produce inaccurate segmentation maps. In the first 100 epochs, we replace PAC^{T} in Fig. 2.2 with

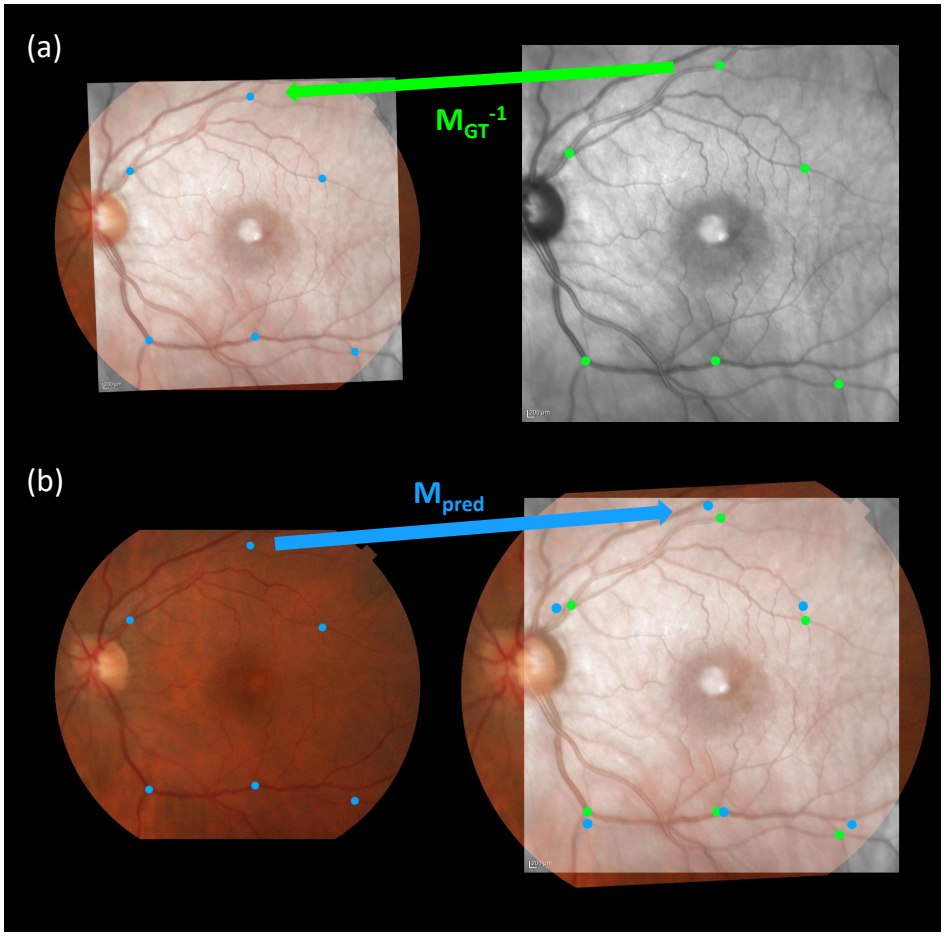


Figure 2.6. Procedure for calculating MAE. (a) Align the target image to the source image using $\mathbf{M}_{\text{GT}}^{-1}$, and record the warped keypoints coordinates. (b) Align the source image to the target image using $\mathbf{M}_{\text{pred}}$, and compute the maximum L2 error on the keypoints to get the MAE.

transposed convolution in the encoder for IR to fit in one GPU. Then, we freeze the weights in the CF encoder and the shared decoder, and train the IR encoder with PAC^{T} for 20 max epochs with learning rate 10^{-4}.

With weights in the encoder and interest point decoder frozen, the SuperPoint descriptor decoder is fine-tuned on 256×256 patches of vessel segmentation maps for our training set. We used $m_p = 1$, $m_n = 0.2$, and $\lambda_d = 250$ for the descriptor loss (2.10), and optimized the network using learning rate 10^{-4}, batch size 2, and 1000 max epochs.

The outlier rejection network is trained with the interest point coordinates after the feature detection and description network is trained on the segmentation maps in our dataset. We set $\lambda_{\text{class}} = 1$, $\lambda_{\text{matrix}} = 0.1$, $\lambda_{\text{Dice}} = 0.1$ for the loss (2.19), and use learning rate 10^{-4}, batch size 32,

and 1000 max epochs. All the coordinates are normalized in $[-1, 1]$, and the transformation matrices are modified accordingly.

Comparison

For comparison, we use the MATLAB code provided by the original authors for [29, 70] and our own implementation in MATLAB for the affine registration part of [30]. Our preliminary method in ICASSP [72] , which utilizes the vessel segmentation [31] and the pretrained SuperPoint network [80], is also evaluated. We then investigate two learning-based methods, CNNGeo [68] and IRR-PWC [93], to choose the better performing method for comparison.

We adopt the pretrained CNNGeo [68] affine model with ResNet101 feature extraction in PyTorch provided by the authors, which is denoted as CNNGeo (pretrained) in Table 2.2. We resize our images to 240×240 for CNNGeo because the input size should be fixed due to the fully connected layers. Note that we also tried to retrain a model with 768×768 input, but the model would not fit in one GPU. We also fine-tune CNNGeo on our training set using the same loss in [68] for 1000 max epochs using Adam optimizer with learning rate 10^{-3} and batch size 16, which achieves better performance than the pretrained model denoted as "CNNGeo (fine-tuned)" in Table 2.2.

Since IRR-PWC [93] uses the pretrained model on FlyingThings3D dataset [104] in PyTorch, we fine-tune the model for our dataset using Adam optimizer with learning rate 10^{-4}, batch size 1, and weight decay 10^{-4} for 100 max epochs. We test IRR-PWC [93] with the pretrained model and the fine-tuned model on our test set, which are denoted as "IRR-PWC (pretrained)" and "IRR-PWC (fine-tuned)" in Table 2.2, respectively.

Table 2.2 shows the registration success rate of the two learning-based methods along with different thresholds for MAE. The success rate of the pretrained models are all 0% for MAE thresholds ranging from 10 to 30 pixels. As we increase the MAE threshold from 10 to 30 pixels, the success rate of fine-tuned IRR-PWC increases from 1.19% to 50.19%, while the success rate of fine-tuned CNNGeo only increases from 0% to 5.53%. IRR-PWC demonstrates better performance than CNNGeo for this task, so we only compare with IRR-PWC for learning-based methods in the

following experiments.

Table 2.2. Comparison between two learning-based methods on JRC CF-IR test set.

Success rate	MAE≤10	MAE≤20	MAE≤30
CNNGeo [68] (pretrained)	0.00%	0.00%	0.40%
CNNGeo [68] (fine-tuned)	0.00%	1.58%	5.53%
IRR-PWC [93] (pretrained)	0.00%	0.00%	0.00%
IRR-PWC [93] (fine-tuned)	1.19%	21.34%	50.19%

Results and Discussion

We compare the proposed method with the conventional (not learning-based) methods and the deep learning-based methods in Table 2.3 for the quantitative result and in Fig. 2.7 for the qualitative result. We used MAE≤10 pixels to determine successful registration in this experiment. As shown in Table 2.3, we first compute the Dice coefficient before registration (by simply padding all image to square shape and resizing to 768×768) for baseline comparison. The average Dice coefficient is expressed as mean ($\pm$ standard deviation).

For conventional methods, SURF-PIIFD-RPM[29] achieves 27.27% success rate and an average of 0.2615 Dice coefficient on the entire test set. The performance improves when excluding bad quality images (both source and target image are "good" or "usable"), but degrades when only considering bad quality images (either source or target image is "bad"). URSIFT-PIIFD-AGMM [70] does perform better than SURF-PIIFD-RPM [29] in some cases in Fig. 2.7, but the overall result is worse than [29] on our dataset, as shown in Table 2.3. Although the Phase-HOG-RANSAC [30] method achieves the highest success rate and Dice coefficient overall among the conventional methods, the 14.00% success rate on bad quality images still indicates very limited robustness.

The pretrained IRR-PWC [93] fails on every image pair, while the fine-tuned IRR-PWC reaches 1.19% success rate and 0.0964 Dice coeffient on average. Since the original input image pairs have large resolution gaps, a wide field of view difference, and large quality variations, applying only a deformable registration method on the original input may not yield accurate results.

In Table 2.3, our proposed method clearly achieves the highest success rate 97.63% and

average Dice coefficient 0.6306 in the test set with 11.07% and 0.0386 improvement upon the best competitor, which is our ICASSP paper [72]. Our proposed method also ranks the highest and reaches 99.51% success rate when excluding bad quality images (both source and target image are good or usable quality). Our method improves the success rate in bad quality images (either source or target image is bad quality) to 90.00% from the previous 50.00% in [72], which significantly improves in robustness thanks to the content-adaptive segmentation method and the fine-tuned SuperPoint network.

Fig. 2.7 shows registration results of different methods for three example pairs in sub-images (1)-(3). For each pair, sub-images (a) and (b) are the input images resized to 768×768. Sub-image (c) shows the checkerboard overlay of the aligned images, where the RGB tiles show the warped source image, and the gray tiles show the target image. Note that the vessels should be continuous across the tiles if the images are well aligned. Sub-image (d) shows the overlay of the aligned segmentation images, where the source segmentation and the target segmentation are assigned to the red channel and the green channel, respectively. The vessels appear in yellow if the segmentation maps perfectly overlap. In the first pair (1), quality of both the source CF image and the target IR image is good. SURF-PIIFD-RPM [29], Phase-HOG-RANSAC [30], and our ICASSP paper [72] succeed, while URSIFT-PIIFD-AGMM [70] and fine-tuned IRR-PWC [93] fail since their MAE is larger than 10 pixels. In the second example (2), which is an image pair with disease and the CF image is blurred, [72] and the proposed method succeeds, while other methods fail. The result of URSIFT-PIIFD-AGMM [70] is close, but the left part is not aligned. For the third image pair (3), where the optic disc in CF image is located near the center of the image, only the proposed method successfully aligned the images. Since the number of keypoint correspondences of [29, 70] is insufficient to calculate an affine transformation matrix (less than 3), their results are set to be the same as before registration.

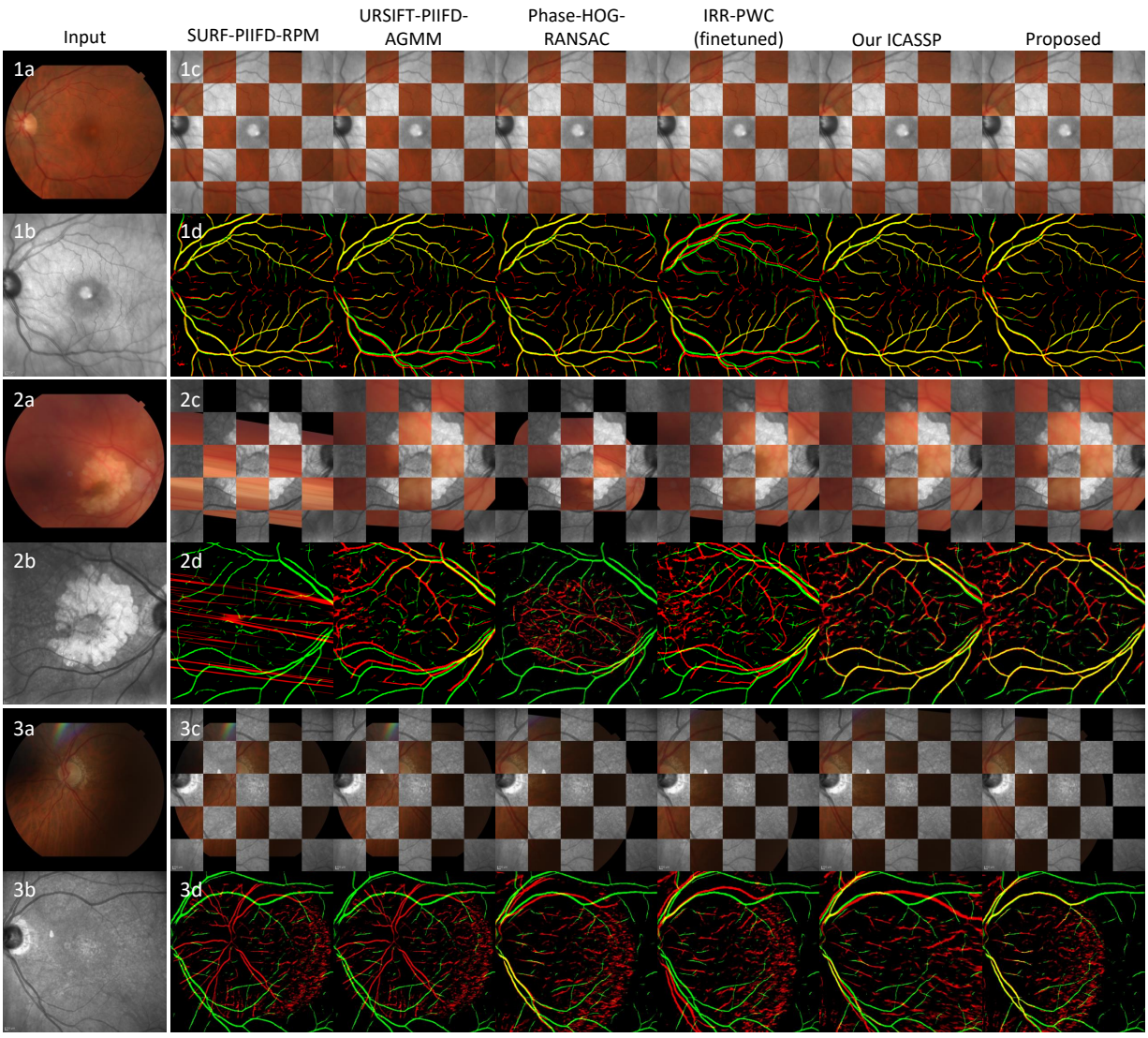


Figure 2.7. Registration results of three example pairs in JRC CF-IR test set using different methods. (a) and (b) show the input image pair resized to 768×768. (c) shows the checkerboard of the aligned original images (RGB tiles: warped source image, gray tiles: target image). (d) shows the vessel segmentation overlay (red: warped source segmentation, green: target segmentation, yellow: overlap).

Table 2.3. Result of different methods with different image qualities on JRC CF-IR test set.

Method	All images (good + usable + bad)		Exclude bad images (good + usable)		Bad images only	
	Success rate	Dice	Success rate	Dice	Success rate	Dice
Before registration	–	0.078 (±0.016)	–	0.080 (±0.015)	–	0.072 (±0.019)
SURF-PIIFD-RPM [29]	27.27% (69/253)	0.262 (±0.245)	33.00 % (67/203)	0.293 (±0.256)	4.00 % (2/50)	0.134 (±0.133)
URSIFT-PIIFD-AGMM [70]	24.90% (63/253)	0.248 (±0.238)	30.05% (61/203)	0.282 (±0.249)	4.00% (2/50)	0.113 (±0.106)
Phase-HOG-RANSAC [30]	40.32% (102/253)	0.331 (±0.262)	46.80% (95/203)	0.372 (±0.262)	14.00% (7/50)	0.162 (±0.180)
IRR-PWC [93] (pretrained)	0.00% (0/253)	0.059 (±0.018)	0.00% (0/203)	0.060 (±0.018)	0.00% (0/50)	0.055 (±0.021)
IRR-PWC [93] (fine-tuned)	1.19% (3/253)	0.096 (±0.060)	1.48% (3/203)	0.102 (±0.064)	0.00% (0/50)	0.073 (±0.033)
Our ICASSP [72]	86.56% (219/253)	0.592 (±0.168)	95.57% (194/203)	0.643 (±0.108)	50.00% (25/50)	0.386 (±0.204)
Proposed	**97.63% (247/253)**	**0.631 (±0.126)**	**99.51% (202/203)**	**0.666 (±0.085)**	**90.00% (45/50)**	**0.485 (±0.154)**
Proposed-ConvSeg	94.86% (240/253)	0.624 (±0.133)	99.01% (201/203)	0.663 (±0.084)	78.00% (39/50)	0.466 (±0.171)
Proposed-RANSAC	84.19% (213/253)	0.589 (±0.154)	91.13% (185/203)	0.633 (±0.106)	56.00% (28/50)	0.411 (±0.190)

Pixel Adaptive Convolution vs Naive Convolution for Vessel Segmentation

To evaluate the benefits of content adaptation, we replace the content-adaptive segmentation network in our image registration framework with a convolutional segmentation network trained in the same way. It is denoted as "Proposed-ConvSeg" in Table 2.3, which achieves 94.86% success rate on all images and 78.00% success rate on bad images. It demonstrates that the convolutional segmentation network is less robust than the content-adaptive vessel segmentation network using PAC in bad quality images.

Fig. 2.8 shows an example pair of images where quality of both the source CF and target IR images are poor. The input images are shown in sub-images (a) and (b), and the mean phase images are shown in sub-images (c) and (d). The mean phase image (b) can clearly represent better vascular details than the original CF image (a). With the guidance of the phase image, the content-adaptive vessel segmentation (f) can also show better details on the right part of the image compared to the convolutional vessel segmentation (e). After applying the fine-tuned SuperPoint and outlier rejection network on the segmentation results, the content-adaptive vessel segmentation achieves successful alignment in sub-image (j), but the convolution version fails in sub-image (i).

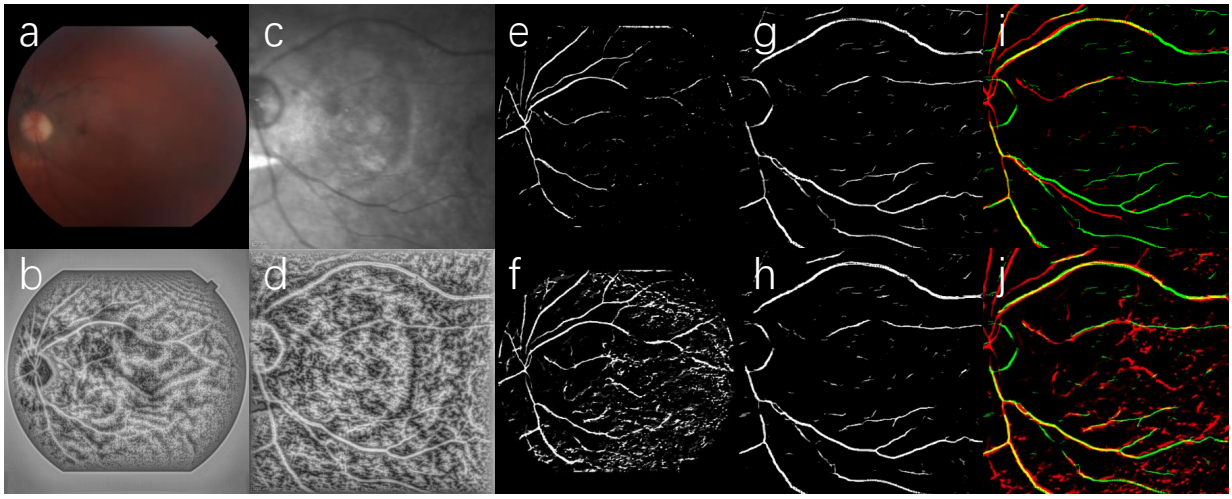


Figure 2.8. Comparison on the PAC and convolution segmentation network. (a), (c): input image pair resized to 768×768. (b), (d): mean phase images. (e), (g): convolution segmentation. (f), (h): PAC segmentation. (i): registration result with convolution segmentation. (j): registration result with PAC segmentation.

Outlier Rejection Network vs RANSAC

We replace the proposed outlier rejection network with RANSAC in our image registration framework to check performance variations, which is denoted as "Proposed-RANSAC" in Table 2.3. Compared to the entire proposed framework, the success rate decreases to 91.13% for good and usable images, and decreases to 56.00% for bad images. It shows that the proposed outlier rejection network leads to significantly more robust performance than RANSAC.

An example is illustrated by Fig. 2.9, where quality of the source image shown in sub-image (a) is bad and quality of the target image shown in sub-image (c) is good. Sub-images (b) and (d) show the keypoints detected by the fine-tuned SuperPoint network, and sub-image (e) shows the mutual nearest neighbor matching result. Sub-images (g) and (h) show the outlier rejection results of RANSAC and the outlier rejection network respectively, where the red lines show the outliers and the green lines show the inliers. RANSAC fails to accurately register the image pair in sub-image (h). However, the outlier rejection network (g) can find more correct inliers than RANSAC (f) such that the alignment result (i) is accurate.

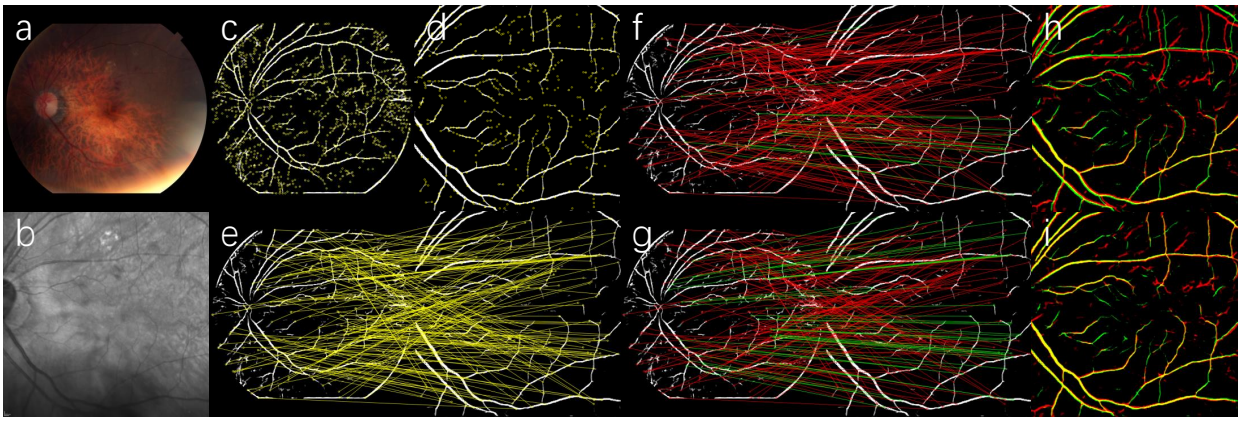


Figure 2.9. Comparison on the outlier rejection network and RANSAC. (a), (b): input image pair resized to 768×768. (c), (d): detected keypoints on the segmentation maps. (e): mutual nearest neighbor matching result. (f): RANSAC matching result (inliers: green, outliers: red). (g): RANSAC alignment result. (h), (i): outlier rejection network results.

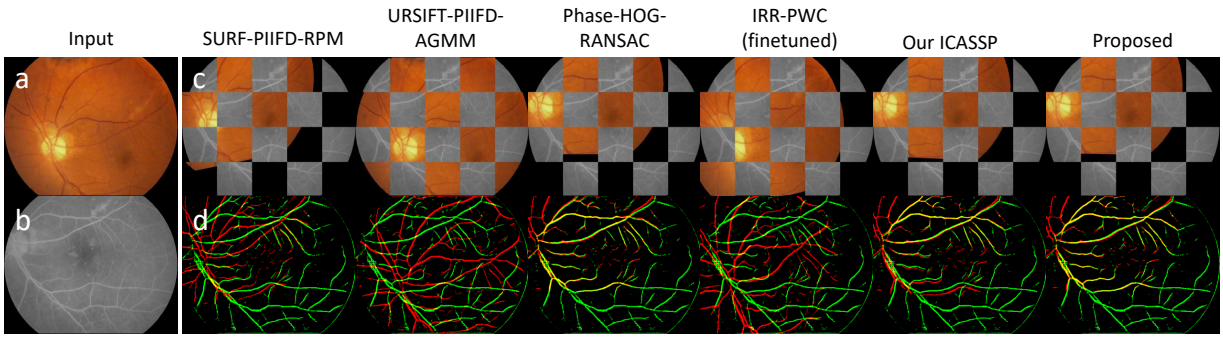


Figure 2.10. Registration results of one challenging pair in CF-FA dataset using different methods. (a) and (b) show the input image pair. (c) shows the checkerboard of the aligned original images (RGB tiles: warped source image, gray tiles: target image). (d) shows the vessel segmentation overlay (red: warped source segmentation, green: target segmentation, yellow: overlap).

2.4.2 CF-FA dataset

We further test the performance of the proposed framework on a public dataset [105], which consists of 59 pairs of color fundus (720×576, RGB) and fluorescein angiography (720×576, grayscale) images. In this dataset, 29 pairs are normal, while the other 30 pairs of images have diabetic retinopathy. The field of view of CF and FA images are similar there are no images with poor quality. We manually labeled 6 pairs of keypoint correspondences to obtain the ground truth transformation matrix for all image pairs. We fine-tuned our proposed framework and other deep learning-based methods [72, 93] on this dataset using similar hyper parameters as in the JRC CF-IR dataset (reducing batch size to 30 for outlier rejection network) for comparison. The 30 pairs with odd indices are used for training, and the other 29 pairs with even indices are used for testing. Due to the small number of images, we did not use a validation set, and simply stop training at the maximum epoch.

Table 2.4 shows the experimental result of different methods on the test set of the CF-FA dataset [105], where the success rate and Dice coefficient are evaluated in the same way as in the JRC CF-IR dataset. Most methods achieve higher success rate compared to the JRC CF-IR dataset overall, due to similar field of view of source and target images and better image quality. Both Phase-HOG-RANSAC [30] and our proposed method achieve 100% success, while our proposed method reached the highest Dice coefficient at 0.6794. When replacing the content-adaptive

Table 2.4. Result of different methods on CF-FA test set.

Method	Success rate	Dice
Before registration	–	0.1197 (±0.0228)
SURF-PIIFD-RPM [29]	82.76% (24/29)	0.5526 (±0.2184)
URSIFT-PIIFD-AGMM [70]	68.97% (20/29)	0.4999 (±0.2402)
Phase-HOG-RANSAC [30]	**100.00% (29/29)**	0.6481 (±0.1004)
IRR-PWC [93] (pretrained)	0.00% (0/29)	0.0984 (±0.0231)
IRR-PWC [93] (fine-tuned)	3.44% (1/29)	0.1977 (±0.0688)
Our ICASSP [72]	96.55% (28/29)	0.6453 (±0.1024)
Proposed	**100.00% (29/29)**	**0.6794 (±0.0873)**
Proposed-ConvSeg	**100.00% (29/29)**	0.6789 (±0.0863)
Proposed-RANSAC	**100.00% (29/29)**	0.6528 (±0.0856)

segmentation in our proposed framework with the convolutional segmentation network, as shown in row "Proposed-ConvSeg", or replacing the outlier rejection network with RANSAC, as shown in row "Proposed-RANSAC", the success rates remain at 100%, but the Dice coefficients decrease slightly.

Fig. 2.10 shows one challenging pair in the CF-FA test set whose the overlapping ratio between source and target images is small. SURF-PIIFD-RPM [29], URSIFT-PIIFD-AGMM [70] and fine-tuned IRR-PWC [93] fail to align this pair, and our ICASSP [72] result also yields MAE larger than 10 pixels. Both Phase-HOG-RANSAC [30] and our proposed method succeed, and our method produces slightly more accurate alignment observed from the segmentation overlay.

The experiment on CF-FA dataset demonstrates that the proposed framework can be easily generalized for other modalities of retinal images via fine-tuning on a small number of images without drastically adjusting hyper parameters. It also shows that the advantages of the proposed framework may be more salient when the dataset includes bad image quality.

2.5 Conclusion

In this chapter, we proposed a content-adaptive weakly supervised deep learning framework for multimodal retinal image registration. The proposed method consists of three neural networks

for vessel segmentation, feature detection and description, and outlier rejection. The content-adaptive nature of the proposed method allows training with weak supervision form ground truth transformation matrices. When compared with recent conventional and learning-based methods, our method achieved the highest success rate and Dice coefficient, and showed significant robustness in bad quality images. In future work, we can concatenate the proposed framework with a locally fine alignment method to form a complete pipeline. The proposed framework could also be further generalized for other modalities including fundus autofluorescence and multicolor imaging.

Chapter 2, in part, is a reprint of the material as it appears in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Y. Wang, J. Zhang, C. An, M. Cavichini, M. Jhingan, M. J. Amador-Patarroyo, C. P. Long, D. G. Bartsch, W. R. Freeman and T. Q. Nguyen, IEEE, 2021, and in IEEE Transactions on Image Processing, Y. Wang, J. Zhang, M. Cavichini, D. G. Bartsch, W. R. Freeman, T. Q. Nguyen, C. An, IEEE, 2021. The dissertation author is the primary author of the two papers.

Chapter 3

Correlation between subjective and objective metrics

Even though the multimodal retinal image registration has been widely studied, the evaluation metrics have not been thoroughly studied and compared. The ophthalmologists use their own subjective grade to assess the accuracy of registration based on the aligned images, while most papers in image processing literature adopt several objective metrics for their fairness and simplicity. This divergence in evaluation metrics may cause performance variations from laboratory to clinical application, and poses potential barriers to improve current registration methods.

In this chapter, we propose a method to mathematically compare the similarity of the subjective grade and the commonly used objective evaluation metrics, and establish an objective evaluation metric that is most correlated with the subjective evaluation of the ophthalmologists. To the best of our knowledge, this work is the first extensive study on various evaluation metrics for multimodal retinal image registration.

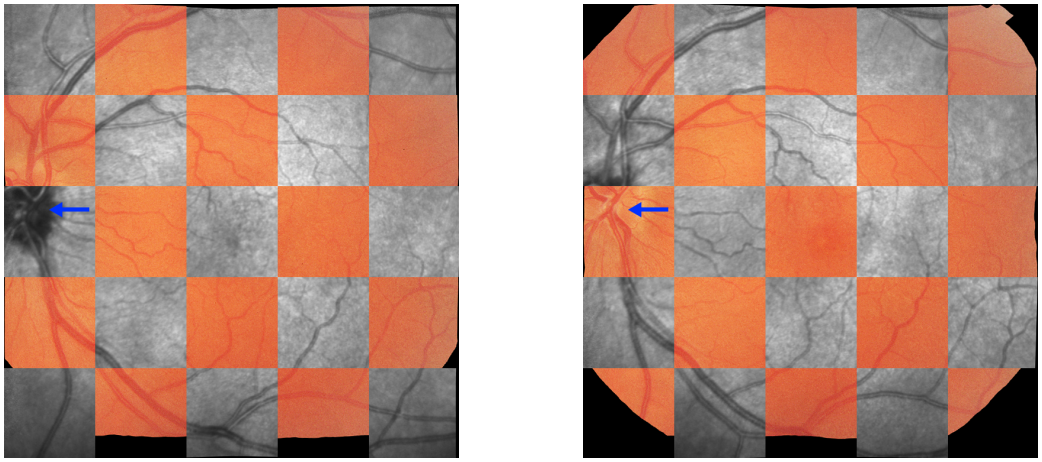
3.1 Subjective metric

Ophthalmologists adopt a subjective grading method, where the aligned multimodal images are analyzed in 5×5 blocks and overlaid in two forms of checkerboard, as shown in Fig. 3.1. The neighboring blocks in the checkerboard image show the registered images between two different modalities. If the images are perfectly aligned, the vessels are expected to be continuous on the edge across the neighboring blocks. In the subjective grading method, a score from 0 to 5 is assigned to

each block in the checkerboard image, characterizing the ratio of overlap in vessels on the edge closest to the optic nerve as specified in Fig. 3.1. The grading criteria are listed in Table 3.1 and the corresponding grade examples are illustrated in Fig. 3.2, where grade 1 and 2 are considered as poor alignment, 3 as reasonable, 4 and 5 as good/excellent matches, and grade 0 is assigned to ungradable blocks due to absence of vessels, or no visible vessels because of noise or out-of-focus. Note that in Figures 3.1 and 3.2, the deformations on the edges of the CF image is caused by the deformable registration method, which will be explained in section 3.4.

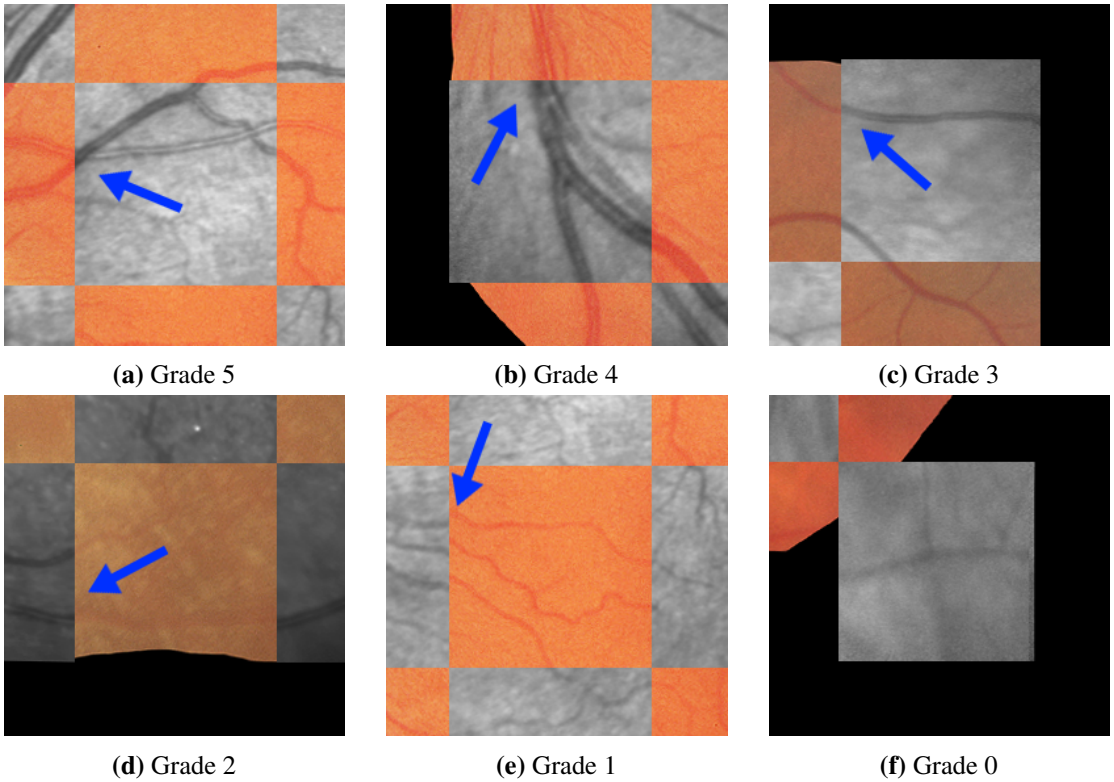
Table 3.1. Grading criteria for the subjective grade.

Grade	Implication (Vessel misalignment at the block boundaries)
5	Perfect alignment.
4	Less than 1/3 vessel width difference in continuity of the vessels.
3	More than 1/3 or and less than 1/2 vessel width difference in continuity between the two vessels.
2	More than 1/2 and less than 1 vessel width difference in continuity between vessels.
1	More than 1 vessel width difference in continuity between vessels.
0	Ungradable due to no visible vessels.



(a) Form 1: Target first **(b)** Form 2: Source first

Figure 3.1. Two forms of checkerboard images are used for subjective grading in (a) and (b), where the blue arrows point to the optic nerve and CF is a source image and IR is a target image for registration.



(a) Grade 5 **(b)** Grade 4 **(c)** Grade 3

(d) Grade 2 **(e)** Grade 1 **(f)** Grade 0

Figure 3.2. Examples of subjective grade. Sub-images (a)-(f) are evaluated from grade 5 to 0. The grade is assigned to the block in the center, and blue arrows point to the vessels are graded based on the Table 3.1 criteria.

3.2 Objective metric

Several objective metrics have been used in image processing literature. In this chapter, we consider Maximum error (MAE), median error (MEE), root mean squared error (RMSE), and percentage of correction keypoints (PCK) for supervised evaluation and structural similarity index (SSIM) and Dice for unsupervised evaluation.

3.2.1 Supervised Evaluation

Supervised evaluation metrics require manually labeled keypoint correspondences between source and target images. Popular metrics include maximum error (MAE) [71, 28, 29, 70], median error (MEE) [71, 28, 29, 70], root mean square error (RMSE) [71, 29, 70, 87], and percentage of correct keypoints (PCK) [68].

Denote keypoint coordinates with $\mathbf{p} = (x, y)^T$ and the set of M pairs of manually labeled

point correspondences with $\mathscr{P} = \{(\mathbf{p}_1, \mathbf{q}_1), ..., (\mathbf{p}_M, \mathbf{q}_M)\}$. MAE (maximum error, different from mean absolute error used in other fields) calculates the maximum L2 norm error on the selected point correspondences

$$\text{MAE} = \max_{(\mathbf{p},\mathbf{q})\in\mathscr{P}} ||F(\mathbf{p}) - \mathbf{q}||, \tag{3.1}$$

where $F(\mathbf{p})$ warps a source keypoint $\mathbf{p}$ towards the target location using the transformation $F(\cdot)$ estimated by the registration algorithm. Using similar notation, MEE calculates the median L2 norm error on the selected points

$$\text{MEE} = \underset{(\mathbf{p},\mathbf{q})\in\mathscr{P}}{\text{median}} ||F(\mathbf{p}) - \mathbf{q}||, \tag{3.2}$$

and RMSE calculates the root mean square error on the selected points

$$\text{RMSE} = \sqrt{\underset{(\mathbf{p},\mathbf{q})\in\mathscr{P}}{\text{mean}} \|F(\mathbf{p}) - \mathbf{q}\|^2}. \tag{3.3}$$

PCK sets a threshold T on the L2 norm to determine whether a pair of keypoints match correctly, and calculates the percentage of correct keypoints

$$\text{PCK} = \frac{|\{(\mathbf{p},\mathbf{q}) \mid ||F(\mathbf{p}) - \mathbf{q}|| < T,\ (\mathbf{p},\mathbf{q}) \in \mathscr{P}\}|}{|\mathscr{P}|} \times 100\%. \tag{3.4}$$

Since the choice of threshold T is task dependent and RMSE with less than 5 pixels is usually considered as success registration [71, 29, 70, 87], the threshold T is set to 5 pixels. To compute these metrics, we need to first manually labeled pairs of keypoint correspondences (generally 6 or more [71, 28, 29, 70, 87]) for all the multimodal images, where the keypoint locations should accurately lie on salient landmarks like vessel bifurcations, and uniformly distributed in the overlapping area. However, it is still difficult for human to select points at pixel-level accuracy and to make the points uniformly distributed in the overlapping area even with the help of user friendly software (e.g. GUI). Besides, labeling point correspondences by hand is very time-consuming for larger datasets with more than hundreds of image pairs.

3.2.2 Unsupervised Evaluation

Unsupervised evaluation metrics, on the contrary, do not require manually labeled keypoint correspondences, and only take the registered images as input.

Denote the aligned images with $\mathscr{I}_1$ and $\mathscr{I}_2 \in \mathbb{R}^{H\times W}$, then the mean square error (MSE) is defined as

$$\text{MSE} = \text{mean}[(\mathscr{I}_1 - \mathscr{I}_2)^2]. \tag{3.5}$$

The structural similarity index (SSIM) [106] is designed to improve MSE, and it is averaged on all the windowed patches $\mathbf{W}_1$ and $\mathbf{W}_2$ from two images $\mathscr{I}_1$ and $\mathscr{I}_2$

$$\text{SSIM}(\mathscr{I}_1, \mathscr{I}_2) = \underset{\mathbf{W}_1,\mathbf{W}_2}{\text{mean}} \frac{(2\mu_1\mu_2 + c_1)(2\sigma_{12} + c_2)}{(\mu_1^2 + \mu_2^2 + c_1)(\sigma_1^2 + \sigma_2^2 + c_2)}, \tag{3.6}$$

where μ_j is the mean value of window $\mathbf{W}_j$ ($j \in \{1,2\}$), and σ_j^2 and σ_{12}^2 are the variance of $\mathbf{W}_j$ and the covariance of $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively. $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ are two variables to stabilize the division with small denominator, with L dynamic range of pixel intensity and $k_1 = 0.01$, $k_2 = 0.03$ by default. SSIM is a value between 0 and 1, and the higher SSIM indicates higher similarity between two images. Since SSIM is only designed for single-channel images image, RGB images need to be first converted to grayscale.

The Dice coefficient is frequently used for evaluating the overlapping region between two segmentation maps [30, 97, 31, 72]. With the aid of vessel segmentation algorithms, the Dice coefficient can also be used to evaluate the accuracy of registration. Let $\mathbf{S}_j$ denote the binary vessel segmentation of image $\mathscr{I}_j$ ($j \in \{1,2\}$), with 1 assigned to vessels and 0 assigned to background, the Dice coefficient for binary segmentation is defined as

$$\text{Dice}(\mathbf{S}_1, \mathbf{S}_2) = \frac{2 \times \sum(\mathbf{S}_1 \odot \mathbf{S}_2)}{\sum \mathbf{S}_1 + \sum \mathbf{S}_2}, \tag{3.7}$$

where $\odot$ denotes element-wise product. The Dice coefficient ranges between 0 and 1, and higher number indicates larger overlap.

The soft Dice coefficient [31] introduces a differentiable counterpart of the binary Dice coefficient for vessel probability maps, and it is defined as

$$\text{Dice}_s(\mathbf{P}_1, \mathbf{P}_2) = \frac{2 \times \sum \text{ele\_min}(\mathbf{P}_1, \mathbf{P}_2)}{\sum \mathbf{P}_1 + \sum \mathbf{P}_2}, \tag{3.8}$$

with $\mathbf{P}_j$ denoting the vessel probability map of $\mathscr{I}_j$ image.

3.3 Correlation evaluation

In order to build a connection between the subjective metric of ophthalmology and the objective metrics of image processing literature, we compute the Pearson correlation coefficient [107] along with various objective evaluation methods to compare their degree of similarity.

3.3.1 Pearson correlation coefficient

The Pearson correlation coefficient (PCC) between random variables X and Y is defined as

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{3.9}$$

where the covariance of X and Y is

$$\text{cov}(X,Y) = \text{E}[(X - \text{E}[X])(Y - \text{E}[Y])], \tag{3.10}$$

and σ_X and σ_Y are the standard deviation of X and Y, respectively. For finite N samples $\{(x_1, y_1), \cdots, (x_N, y_N)\}$, PCC can be estimated by

$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}. \tag{3.11}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ denotes the sample mean, and similarly for $\bar{y}$. The PCC value ρ ranges between $[-1, 1]$, where higher absolute PCC value $|\rho|$ implies higher linear dependency. Positive ρ value

implies that Y increases as X increases, while negative value implies that Y decreases as X increases. For example, the correlation between MAE and the subjective grade should be negative, while the correlation between PCK and the subjective grade should be positive. Based on Cohen's interpretation [108], an absolute PCC value of 0.1 is small, 0.3 is medium, and 0.5 is large. An important property of the PCC value is that it is invariant to separate linear transforms in X and Y

$$\rho_{X,Y} = \rho_{(aX+b),(cY+d)}, \text{ where } ac > 0,\ a,b,c,d \in \mathbb{R}, \tag{3.12}$$

which means that PCC can still be applied even when the scale and range are different for different criteria.

3.3.2 Confidence interval

In order to test the certainty of the estimated PCC r in eq (3.11), we further calculate its confidence interval to indicate the possible range of true PCC ρ values in eq (3.9). The confidence interval for the population's PCC ρ is can be computed with three steps [109]. Firstly, a z-score is computed applying the Fisher transformation $F(\cdot)$

$$z = F(r) = \frac{1}{2}\ln\frac{1+r}{1-r} = \operatorname{arctanh}(r). \tag{3.13}$$

Secondly, given a significant level α which is typically set to 95%, the critical z-score $z_{\alpha/2}$ for a two-tail test can be obtained from a look-up table. Then the confidence interval for $F(\rho)$ would be

$$F(\rho) \in \left[F(r) - z_{\alpha/2}\mathrm{SE}, F(r) + z_{\alpha/2}\mathrm{SE}\right], \tag{3.14}$$

where the standard error SE is $\frac{1}{\sqrt{n-3}}$. Finally, the confidence interval for r is converted back using the inverse Fisher transformation

$$F(\rho) \in \left[F^{-1}\left(F(r) - z_{\alpha/2}\mathrm{SE}\right), F^{-1}\left(F(r) + z_{\alpha/2}\mathrm{SE}\right)\right], \tag{3.15}$$

where

$$F^{-1}(z) = \frac{e^{2z}-1}{e^{2z}+1} = \tanh(z). \tag{3.16}$$

3.4 Experimental result

We derive the Pearson correlation between various objective metrics and the subjective grade on the multimodal retinal image registration result. We apply two automatic deformable registration methods to obtain aligned images: The first one is a conventional method, MIND [110], and the second is a learning-based method [31]. As the deformable registration methods are designed for small misalignment, source and target images are aligned with ground truth and then they are deformed randomly [97, 31], as shown in Fig. 3.3. Finally, we apply two registration methods [31, 110] to generate the aligned images. Ophthalmologists grade them for the subjective metric, and we derive the objective metric with MAE, MEE, RMSE, PCK, SSIM and Dice coefficients. Note that here, we use two automatic deformable registration methods [31] and [110] not because of comparing performance between two methods but because of providing consistency of correlation between the subjective and objective metrics for both the conventional and learning methods.

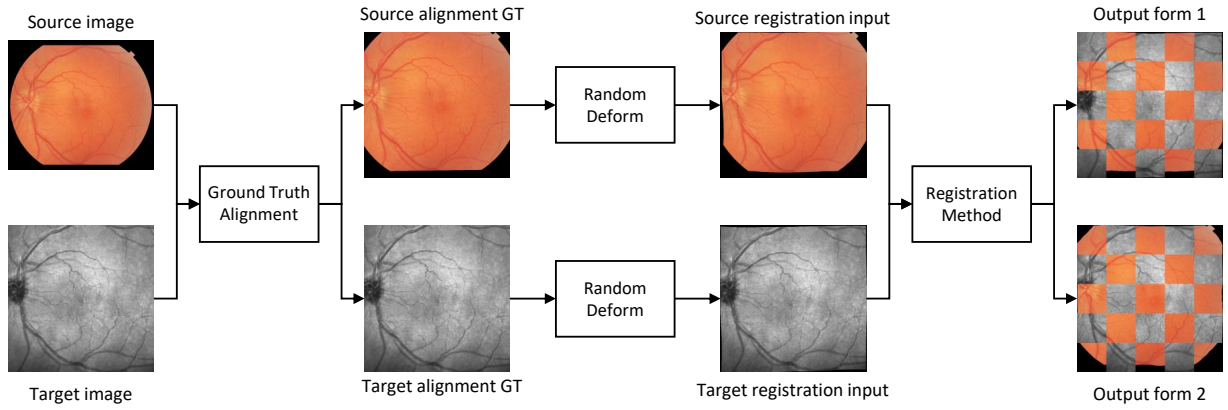


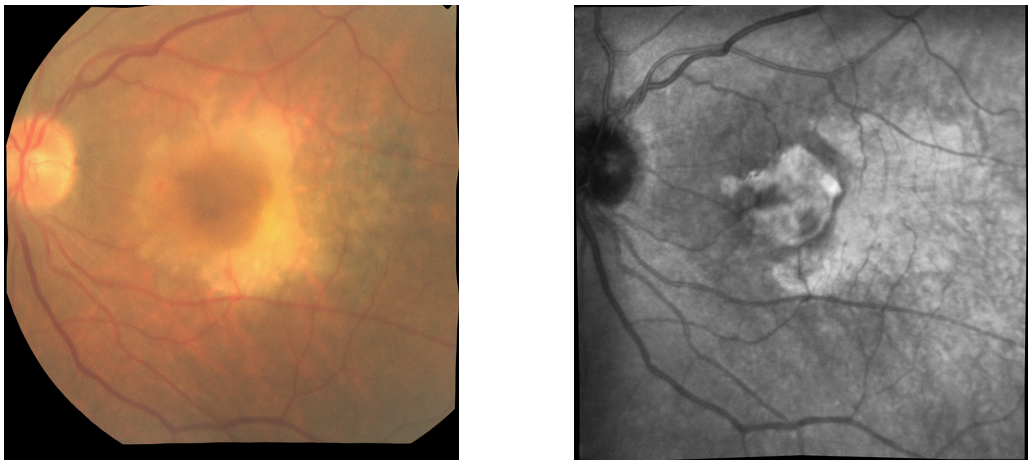
Figure 3.3. Input generation and deformable registration pipeline where registration methods by Zhang et al. [31] and Heinrich et al. [110] are used.

3.4.1 Dataset and ground truth

The dataset consists of 109 pairs of Color Fundus images (TRC-50DX color fundus images, Topcon) and infrared Scanning Laser Ophthalmoscope images (Heidelberg Engineering Spectralis SLO). The dataset includes a variety of pathologies including hemorrhages, diabetes, and macular degeneration.

In this experiment, multimodal images include 3 different types of content change. Firstly, the pathology has different appearance in the two imaging modalities, as illustrated in the example in Fig. 3.4. Secondly, the images are not necessarily taken at the same time, which also changes appearance as the disease progresses. Thirdly, images are distorted by artifacts such as out-of-focus, reflection, and over or under exposed regions.

The source CF images are 24-bit RGB and their resolution is 3000×2672, and the target SLO images are 8-bit grayscale and resolution of the images is 768×768 or 1536×1536. They are both padded to square shape and resized to 768×768 before registration, and the ground truth transformation matrices are derived by manually selected correspondences from the source to target images. We use the SuperPoint network [80] to detect keypoints on the vessel segmentation of [31] and select all possible pairs of point correspondences to compute MAE, MEE, RMSE, and PCK.



(a) CF image **(b)** IR image

Figure 3.4. Example image pair with disease (aligned by [31])

3.4.2 Experiment setting

As shown in Fig. 3.3, the input is generated by randomly deforming the image pair aligned by the ground truth transformation matrix, where the random deformation field is upsampled from a $4 \times 4 \times 2$ matrix that follows normal distribution with 0 mean and standard deviation of 5 pixels. The input pair is then registered using each of the two registration methods to generate the output pair, where MIND [110] is implemented in MATLAB, and [31] is implemented in PyTorch.

Outputs of two registration methods are graded independently by 2 retina specialists and 3 medical students. All pairs of aligned images are graded two times based on two forms of checkerboard overlay shown in Fig. 3.1. We take the average of the two grades assigned to the same block if both of them are non-zero, take the non-zero grade if only one grade is zero, and keep the grade zero if both grades are zero. To verify the reliability of the grades, each of the 5 graders first grades the same subset of 10 images, and applies SPSS [111] to calculate an average intraclass correlation coefficient (ICC). We obtain that the average ICC is 0.903, which indicates excellent similarity among graders.

In the comparison, each block is treated as an individual sample, as illustrated in Fig. 3.5. MSE, SSIM, and Dice coefficients are computed on the image patch within each block, where the black area in CF image is excluded. MAE, MEE, RMSE, and PCK are calculated on all correspondences within the block, as shown in Fig. 3.6. Especially, blocks with no valid correspondences or marked as ungradable (grade 0) are excluded when calculating the correlation coefficient, otherwise they may affect majorly the correlation coefficients.

We set the threshold T in eq. (3.4) at 5 pixels for PCK, which yields the maximum correlation with grade at 0.1539, as shown in Fig. 3.7. MSE and SSIM are computed on the aligned images after converting to grayscale. Two segmentation methods [35] and [31] are tested for the binary and soft Dice coefficients where the optimal thresholds of the binary Dice coefficient for two methods are set to 0 and 0.5, respectively, which yields largest correlation with the grade.

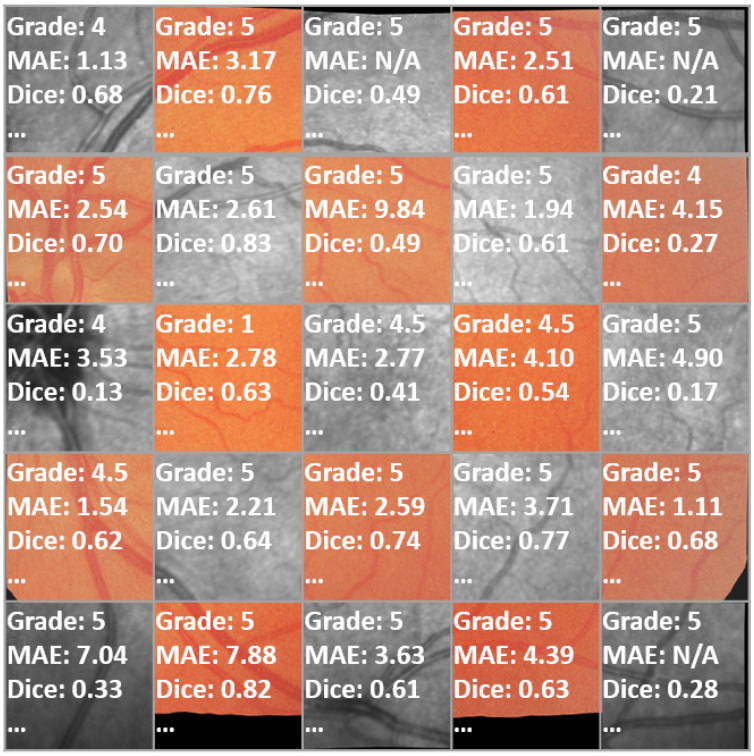
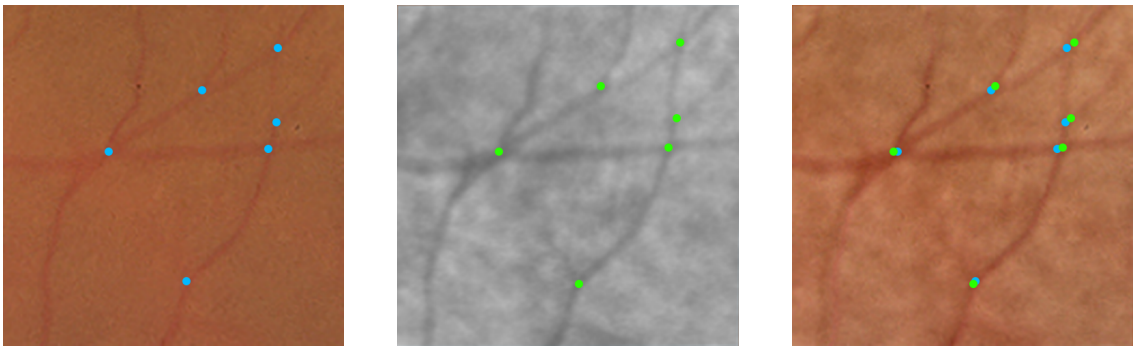


Figure 3.5. Subjective and objective metrics on each block in an example image.

3.4.3 Results

Fig. 3.8 shows the correlations between the subjective grade and various objective metrics and their 95% confidence intervals, which are evaluated on each block or image pair. The soft Dice coefficient using segmentation [31] achieves the highest correlation with the subjective grade at 0.35, which can be considered moderate based on Cohen's interpretation [108], and the binary Dice coefficient [31] shows slightly lower correlation. The binary and soft Dice coefficient in [35] also demonstrate lower correlation with the subjective grade. This result coincides with the fact that the vessel segmentation of [31] is more accurate than that of [35]. PCK and SSIM show positive correlation around 0.1, while the error metrics such as MAE, MEE, RMSE, and MSE show negative correlation around -0.1. The main reason that the binary and soft Dice coefficient show the highest correlation with subjective grade may lie in their common emphasis on retinal vessels. The Dice coefficient calculates the overlapping of vessel segmentation, and the subjective grade also depends on the continuity of vessels. Meanwhile, the other metrics do not explicitly depend on vessels,



(a) Source keypoints **(b)** Target keypoints **(c)** Correspondences

Figure 3.6. Keypoint correspondences in an example block.

which leads to lower correlation with the subjective grade.

In order to evaluate the intrinsic correlation between different metrics, we calculate the absolute PCC value between each pair of metrics shown in Fig. 3.9. We observe that the correlation between the supervised metrics are very high, because they all operate on the selected keypoint correspondences. MSE and SSIM also have very large correlation, as they both take the grayscale images as input. The binary and soft Dice coefficient with the segmentation [31] are nearly perfectly correlated, as the segmentation map of [31] is already close to binary. Meanwhile, the binary and soft Dice coefficient with the segmentation [35] are different since the vessel segmentation is similar to the probability map.

We further investigate the influence of different registration method. In the following experiment, correlation is calculated separately using the two registration methods [31, 110]. As shown in Fig. 3.10, the soft Dice coefficient with segmentation [31] still shows the highest absolute correlation in both methods. Specifically, the soft Dice coefficient with segmentation [31] has higher correlation using only registration method [110], where the grade distribution in Fig. 3.11, better covers both higher and lower grades. This experiment demonstrates that adopting different registration methods will lead to similar conclusion.

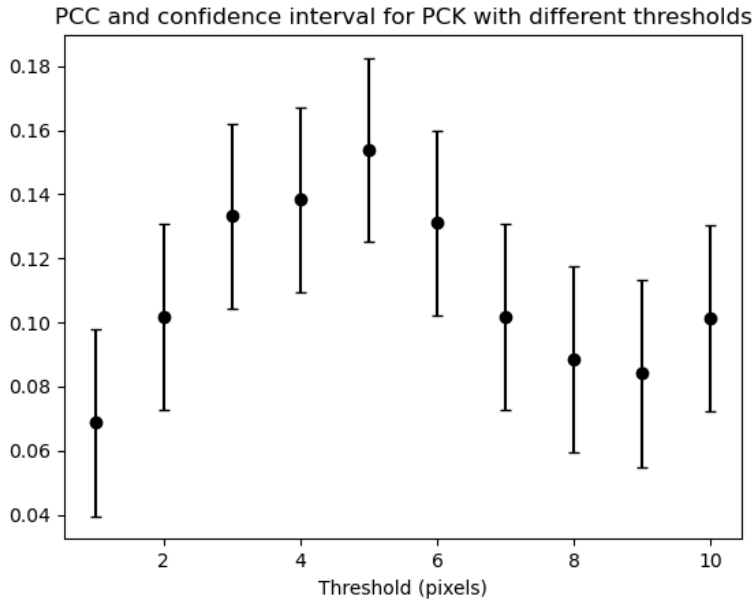


Figure 3.7. Pearson correlation coefficient between subjective grade and PCK with different thresholds with 95% confidence interval.

3.5 Conclusion

In this chapter, we presented a comprehensive overview of the existing evaluation metrics for multimodal retinal image registration, and compared the Pearson correlation coefficient between the ophthalmologists' subjective grade and several commonly used objective evaluation metrics. We found that the soft Dice coefficient with the segmentation method in [31] achieved the highest correlation with the subjective grades compared to many commonly used keypoint-supervised metrics. This study established an objective metric, i.e. the Dice coefficient, that is highly correlated with the subjective evaluation of the ophthalmologists. The experimental results would build a connection between ophthalmology and image processing literature, and the findings may provide a good insight for researchers who investigate retinal image registration, retinal image segmentation and image domain transformation.

Chapter 3, in full, is a reprint of the material as it appears in IEEE Access, Y. Wang, J. Zhang, M. Cavichini, D. G. Bartsch, W. R. Freeman, T. Q. Nguyen, C. An, IEEE, 2021. The thesis author is the primary author of the this paper.

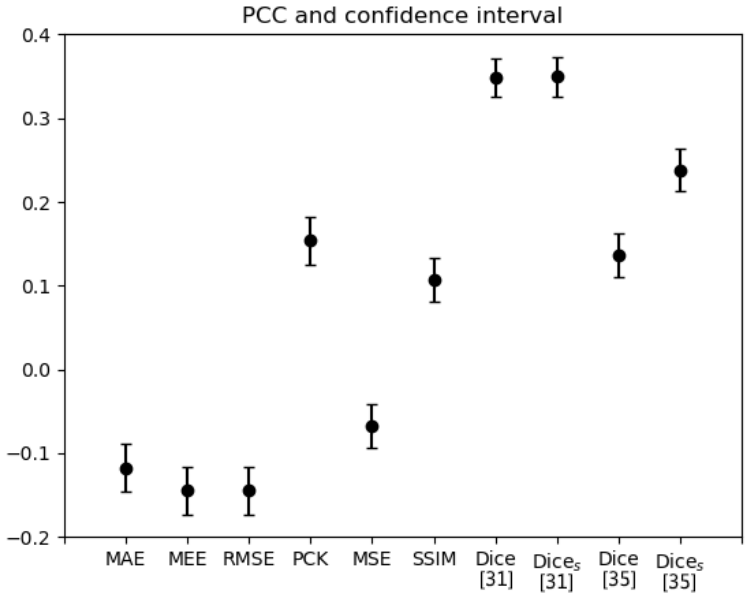


Figure 3.8. Pearson correlation coefficient between subjective grade and objective metrics with 95% confidence interval using two registration methods (Zhang et al. [31] and Heinrich et al. [110]).

Figure 3.9. Absolute value of Pearson correlation coefficient between different metrics.

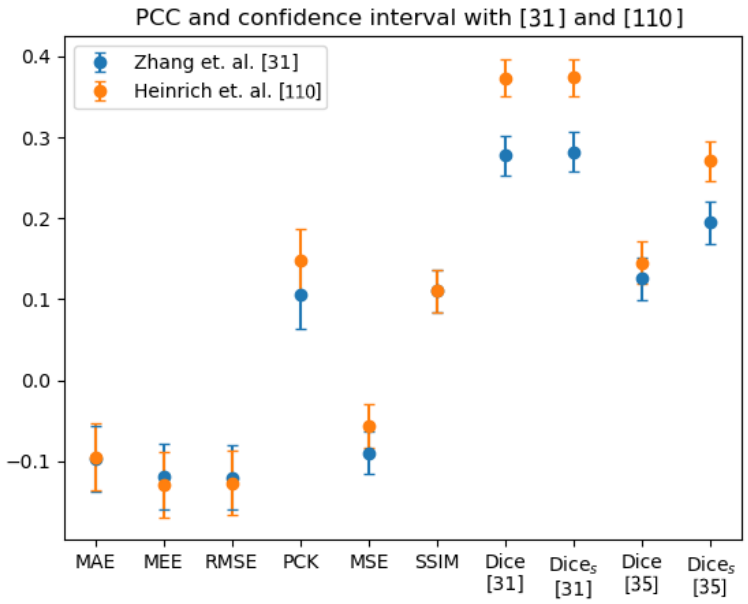


Figure 3.10. Pearson correlation coefficient between subjective grade and objective metrics with 95% confidence interval using registration method Zhang et. al. [31] or Heinrich et. al. [110] only.

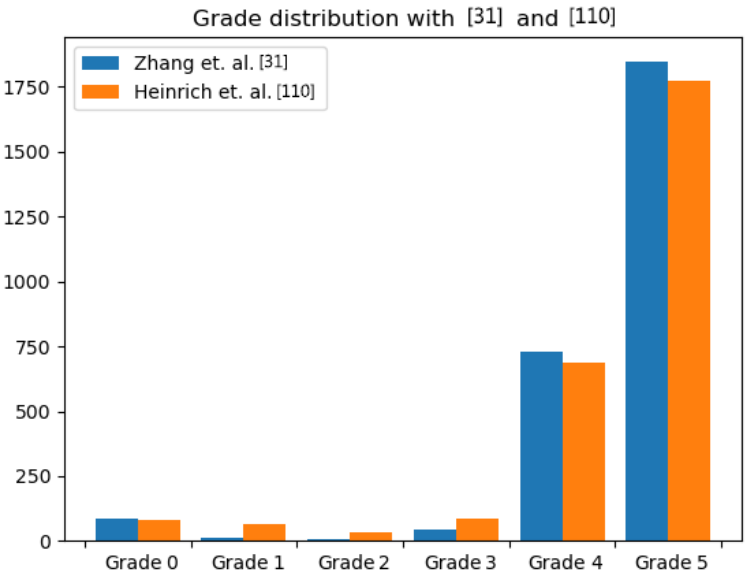


Figure 3.11. Distribution of subjective grade using registration method Zhang et. al. [31] or Heinrich et. al. [110].

Chapter 4

OCT motion correction

4.1 Introduction

Optical Coherence Tomography (OCT) is a non-invasive imaging technique that visualizes cross-sectional images of biological tissues at micrometer-resolution [112]. The impact of OCT in retinal imaging is so significant in opthalmology, that OCT has become the standard of care for diagnosing and monitoring most retinal diseases [2], including age-related macular degeneration (AMD), diabetic macular edema (DME), glaucoma, and so on.

The imaging principle of OCT is based on low-coherence interferometry. The object is probed with low-coherent infrared light, and the depth of backscattered light along the beam axis is measured by interference. The interferogram intensities represent 1D depth (A-scan, Z axis of Fig. 4.1) from the backscattering. 2D cross-sectional images (B-scan, XZ plane of Fig. 4.1) are acquired in a sequence by moving the infrared beam through the object in a raster-scanning pattern. Finally, a 3D volume can be formed by stacking the B-scans (XZ planes) to the Y axis, as illustrated in Fig. 4.1. The direction for B-scan acquisition (X axis of Fig. 4.1) is called the *fast scanning axis*, while the direction for stacking B-scans (Y axis of Fig. 4.1) is called the *slow scanning axis*, and the plane spanned by the other two axes is called the *coronal* or *en-face* plane.

Motion correction is one of the major challenges in OCT imaging, as motion artifacts would not only influence the visualization of the volumetric data, but also reduce the reliability of retinal biomarkers [46]. Moreover, they may increase the difficulty of downstream tasks including disease

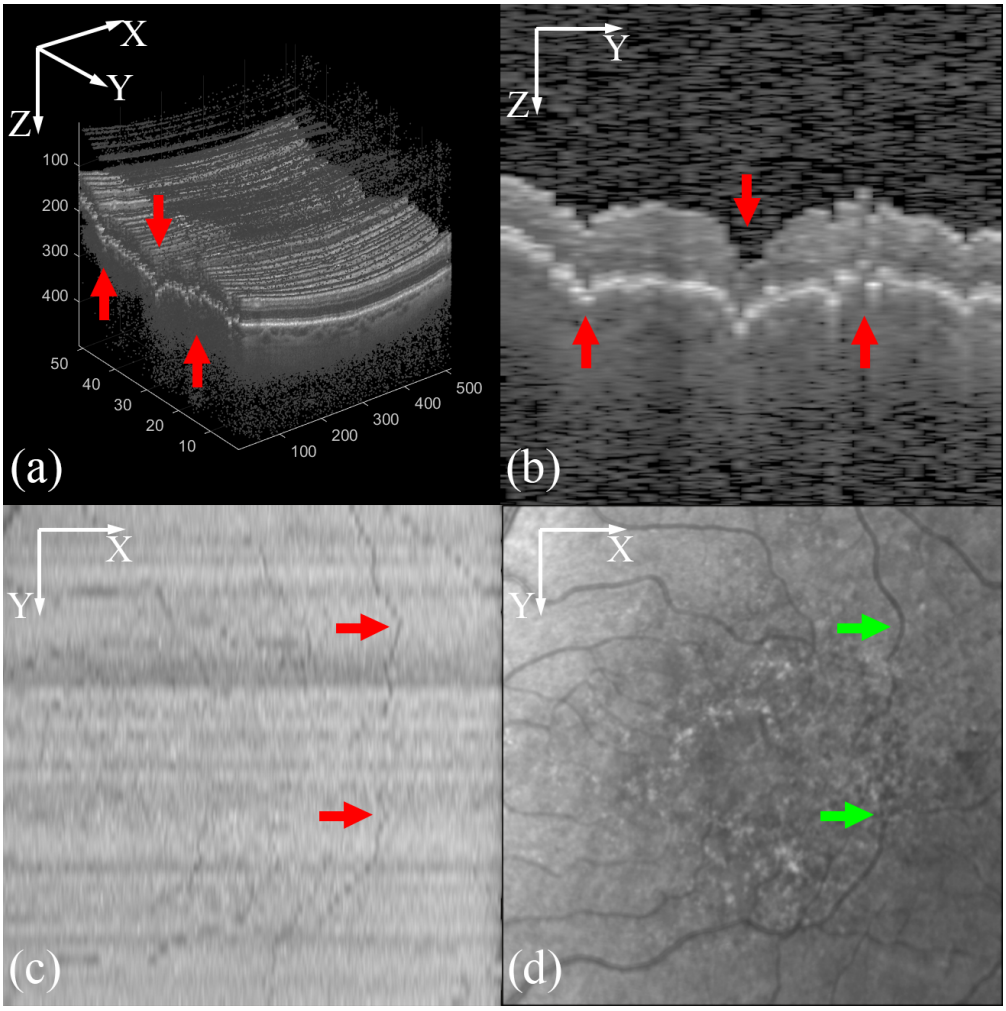


Figure 4.1. Axial and coronal motion artifacts in OCT. (a) The axial motion artifacts in 3D OCT volume indicated with red arrows, (b) cross-sectional B-scan (YZ plane) with motion artifacts, (c) en-face C-scan (XY plane) with motion artifacts, (d) IR en-face image reference with true vessel shapes indicated by green arrows.

classification and segmentation, layer segmentation and OCT-Angiography (OCT-A) imaging. Motion artifacts in OCT can be caused by head motion, respiration, vascular pulsation, and involuntary fixational eye movements. Specifically, even when the patient is instructed to stay still and fixate on an object, the eye may still carry out microsaccades, drifts, and tremors with various frequency and magnitude [46]. These involuntary motions lead to axial and coronal distortions, shown in sub-images 4.1(b) and (c) respectively, where the motion artifacts are indicated by red arrows. The axial motion introduces discontinuity in the cross-sectional B-scan as in sub-image (b), while coronal motion causes distortion of vessels in the en-face plane shown in sub-image in Fig. 4.1(c)

compared with the reference IR image in sub-image (d).

Existing literature on OCT motion correction can be categorized into *prospective* and *retrospective* approaches [46]. Prospective approaches include hardware eye-tracking systems that detect and compensate for motions during image acquisition [51, 52]. Although hardware-based methods can usually achieve more accurate results, they are not available for every OCT device and their solution cannot eliminate all the motion artifacts. Retrospective approaches are applied after image acquisition and most are software-based methods. Many successful software methods require more than one OCT volume [47, 53] or multimodal images as a reference, which introduces extra burden for clinical tests. Other methods based on a single OCT volume tend to remove the curvature of the retina generating overly smoothed result [54, 55].

In this chapter, we propose a deep learning method that utilizes fully convolutional neural networks to correct both axial and coronal motion artifacts in OCT with a single input volume. The axial motion correction network is able to recover the overall curvature of the retina, which compensates for the remaining error in eye-tracking hardware. To the best of our knowledge, the proposed axial motion correction network is the first application of deep learning in the OCT motion correction problem, and improvements upon conventional methods are demonstrated by experimental results. We also present the preliminary results of the coronal motion correction network and discuss possible improvements for future work.

4.2 Related work

Although artifacts caused by involuntary eye motion is a fundamental problem in retinal OCT, it has not been widely researched according to existing literature. Existing literature in OCT motion correction have been extensively reviewed and summarized in two papers published in 2017 [113] and 2019 [46]. It should be noted that this paper focuses on the inter-B-scan motion correction problem in OCT. The time-domain OCT tracking problem [114] is not considered in this paper, since the problem assumes motion-free 3D OCT volume and performs motion estimation in time-domain. OCT-A domain motion artifacts correction [115] is also beyond the scope of this

paper, as motion artifacts in OCT-A appear as a temporal-domain misalignment such that brighter and noisy scans, while motion artifacts in OCT appear as a spatial-domain misalignment such that unaligned raster scans.

Most works correct axial and coronal movement by treating fast B-scans as artifact-free rigid bodies, since the acquisition speed of a fast B-scans is faster than that of the expected eye motion using modern OCT devices [46]. Axial movement is observed to be more significant than coronal movement in magnitude [47], and the axial resolution is also often higher than coronal resolution [47, 53]. Coronal artifacts are caused by eye movement in the 2D en-face plane. The X component of such motion, also called *in-plane* motion, is parallel to the X-Z plane of B-scans and can be observed by discontinuities in retinal vessels, which are prominent features in the en-face plane. The Y motion along the slow scanning axis, also called *out-of-plane* motion, is most difficult to quantify [113]. Negative or positive displacement to the Y axis can cause repeated B-scans of the same region or larger gaps between neighboring B-scans.

Most existing approaches can be divided in two major categories: prospective and retrospective approaches [46]. Prospective approaches often include active eye-tracking hardwares mounted on OCT devices [51, 52] and usually produce more accurate alignment results [113]. There are also approaches that depend on specially designed scanning patterns [116] or signal acquisition techniques [117] to obtain an artifact-free OCT volume. Nevertheless, they are difficult to implement in existing OCT systems and cannot correct eye movements in conventional OCT scans. The retrospective approaches on the other hand are software-based solutions. Many motion correction algorithms require more than one OCT volume, either repeated in the same direction [118, 119], or both horizontal and vertical scans in orthogonal directions [47, 53]. Potsaid et al. applied orthogonal OCT volumes in both horizontal and vertical directions that corrects the axial motion [47]. Gibson et al. proposed an axial motion correction algorithm based on optic nerve head segmentation surface that requires parallel OCT volumes scanned in the same direction [118]. Niemeijer et al. proposed a graph-based method to register multiple OCT volumes and find the optimal translation for each A-scan that can correct both X and Z motion [119]. Wu et al. proposed

a registration method for 3D OCT volumes based on Coherent Point Drift [40], which includes a motion correction algorithm that detects and deforms 3D vessel center lines in source and target volumes. The method proposed by Kraus et al. [53, 120] registers two or more OCT volumes in orthogonal directions, which sequentially corrects axial and coronal motion in two stages. It has been widely adopted as a standard pre-processing algorithm in OCT-A imaging [46]. However, these methods need to capture multiple OCT volumes, such that they increase (double) the time required for clinical examinations and impose additional burden on limited medical resources.

Other methods for estimating eye motion based on a single OCT volume are proposed to save imaging time. The method proposed by Antony et al. [54] utilizes segmentation of the retinal pigment epithelium (RPE) layer and thin-plate spline fitting. A major drawback of the method is that it flattens the RPE layer into a plane, which leads to artifacts for diseases that manifest in the RPE layer. It is also undesirable to observe diseases including myopic [121, 122] as it removes the curvature of the retina. Xu et al. [55] proposed a particle filter method to correct axial and X directional motion in optic nerve head (ONH) centered OCT volumes. The method also results in flattened RPE surface, and is only validated on synthetic motions within 2-10 pixel range. The algorithm of Montuoro et al. [48] corrects axial motion by smoothing the RPE segmentation with a local symmetry assumption, and accounts for X directional coronal motion based on phase shift in the Fourier domain. It can recover the retinal curvature to some extent, but the symmetry assumption does not always hold for retina with diseases. Fu et al. [123] proposed a method to correct both axial and X directional motion based on saliency detection, but the authors only tested the performance of the X motion correction using synthetic data with X motion smaller than 15 pixels. Abdolmanafi et al. [124] proposed a hybrid method for intracoronary OCT. The method combines features from pre-trained AlexNet with conventional searching based on cosine similarity. However, the method is not end-to-end optimized as the AlexNet features are not directly trained for motion correction performance. Besides, the inference time of their method takes 119 minutes for each OCT volume, which is difficult to be applied in practice.

In our previous work [125], we applied an end-to-end deep learning algorithm to correct the

axial motion. It improved correction accuracy while preserving retinal curvature. In this paper, we propose deep learning networks to jointly correct both coronal and axial motions. There are many advantages of the proposed end-to-end neural networks compared to conventional [54, 55, 53] or hybrid [124] methods. The proposed method can be end-to-end optimized, can be generalized to various diseases and resolutions without tuning, and it can reduce the computational time from nearly 2 hours [124] to 0.29 seconds when accelerated by a GPU. We include a comprehensive comparison with several methods using different metrics for OCT volumes with analysis of various diseases and different resolutions. The preservation of curvature is explicitly evaluated by the curvature and distortion coefficient. Ablation studies of segmentation input, post-processing, and stand-alone coronal motion correction network are included. We also add qualitative comparison of the layer segmentation and vessel segmentation using different motion correction methods.

4.3 Axial motion correction network

We first apply axial motion correction to eliminate large Z displacement before correcting coronal motion, because the axial motion is more significant compared with coronal motion in retinal OCT. It has been observed that the axial motion is larger in micrometers compared with coronal motion [47, 53], and the fact that the axial direction has higher resolution in most OCT systems [47, 126, 127] makes the axial artifacts more dominant.

4.3.1 Network architecture

We present a modified U-Net [66] with residual blocks to predict a displacement map for a single OCT volumetric input. The proposed network takes any number of stacked B-scans due to the fully convolutional architecture. The proposed baseline method of Fig. 4.2 operates on a single OCT volumetric scan $\mathbf{V} \in \mathbb{R}^{H \times W \times N}$ where W and H are the width and height of each B-scan and N is the number of B-scans. Z axis of the input OCT volume is treated as channels (H), while X and Y axes are considered as spatial dimensions ($W \times N$). The network outputs a displacement map $\mathbf{D}_z \in \mathbb{R}^{W \times N}$ where each pixel contains a displacement value to Z axis. Negative displacement shifts

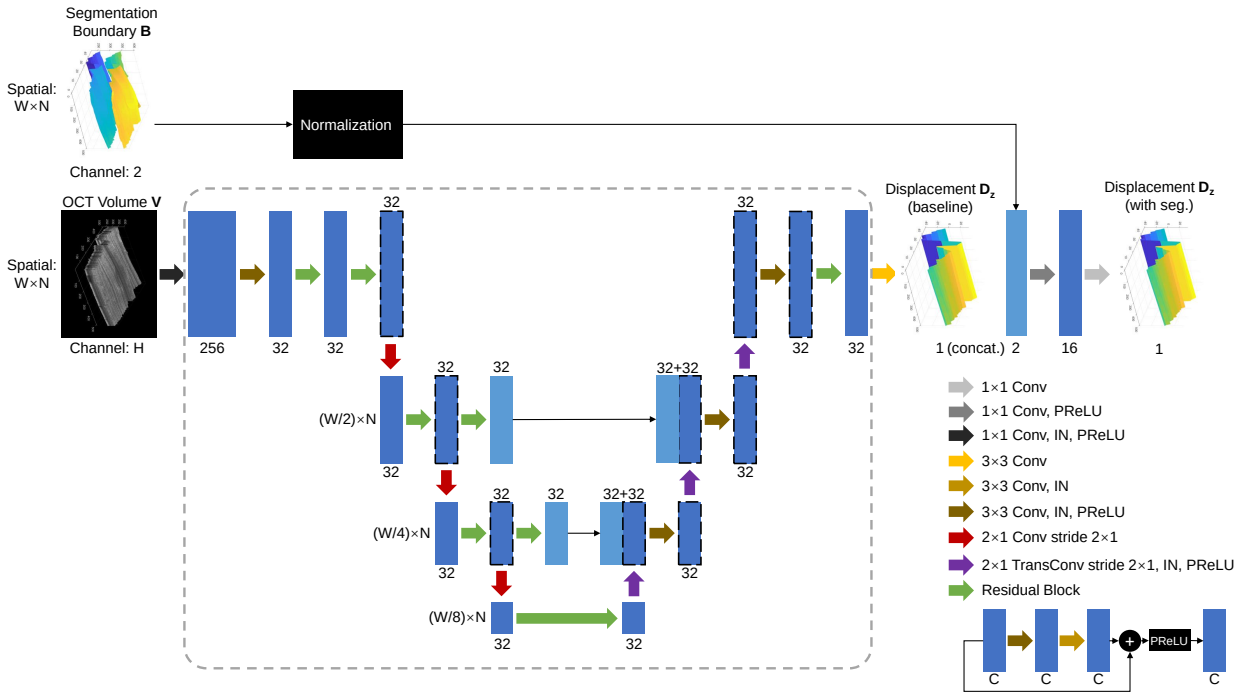


Figure 4.2. Architecture of the proposed OCT motion correction network to predict an axial displacement map. The baseline network is circled in dashed lines, and the network with segmentation concatenates the baseline output with the normalized segmentation boundary to enhance the final displacement prediction.

the A-scan upwards and positive displacement shifts it downwards. The magnitude of displacement denotes the number of pixels to be shifted and it is divided by a normalization factor Z_{norm} , which is a tunable hyper-parameter to scale predicted Z displacement, for better numerical stability. Finally, the motion corrected OCT volume $\mathbf{V}_{\text{dz}}$ is derived for a given displacement $\mathbf{D}_z$ in floating point value as

$$\mathbf{V}_{\text{dz}}(z,x,y) = \mathbf{V}(z - \text{int}(Z_{\text{norm}}\mathbf{D}_z(x,y)),\ x,\ y), \tag{4.1}$$

where $(x,y,z) \in [0,W-1] \times [0,N-1] \times [0,H-1]$ and $\text{int}(\cdot)$ denotes integer conversion.

As illustrated in Fig. 4.2, the network includes 4 resolution levels similar to U-Net. 1×1 convolution is applied at the first layer to compress the number of channels, and it is followed by 3×3 convolutions for further processing. Instance normalization (IN) is applied after convolutions in order to normalize over the spatial dimensions without being influenced by other volumes in the same batch. The skip connection is removed at the original resolution to enhance smoothness of prediction and reduces memory consumption. For the three downsampling blocks denoted by red

arrows, 2×1 convolution with stride 2×1 is adopted to downsample the X dimension by 2, while keeping the resolution on the Y dimension unchanged, since the number of B-scans N in our dataset is significantly smaller than the width of B-scans W. Similarly, 2×1 transposed convolution with stride 2×1 is used to upsample X dimension by 2 in three upsampling blocks which are indicated by purple arrows. Differently from the original U-Net, the input features on the same resolution are processed by residual blocks before concatenation with the upsampled ones. Dropouts are applied at blue blocks with black dashed contour in Fig. 4.2 to prevent overfitting during training.

We also propose an enhanced version of the baseline architecture by including the segmentation of the inner limiting membrane (ILM) and the retinal pigment epithelium (RPE) layer. As shown in Fig. 4.2, we first normalize two segmentation boundaries and concatenate them with the output of the network, and then apply two additional layers with 1×1 convolution to get the displacement prediction. The boundary normalization is computed by the following steps. We denote the two segmentation boundaries with $\mathbf{B} \in \mathbb{R}^{2 \times W \times N}$, where $\mathbf{B}(0,x,y)$ and $\mathbf{B}(1,x,y)$ entries represent Z coordinates of ILM and RPE layers at pixel (x,y), respectively. The overall retinal tilt $\mathbf{T} \in \mathbb{R}^{2 \times W \times N}$ is first computed by

$$\mathbf{T}(z,x,y) = \mathbf{B}(z,x,0) + \frac{\mathbf{B}(z,x,N-1) - \mathbf{B}(z,x,0)}{N-1} y, \tag{4.2}$$

where $z \in \{0,1\}$ and $(x,y) \in [0, W-1] \times [0, N-1]$. Then, the normalized boundaries $\mathbf{B}'$ can be obtained by

$$\mathbf{B}' = (\mathbf{T} - \mathbf{B}) / Z_{\text{norm}}. \tag{4.3}$$

4.3.2 Ground truth acquisition

In order to obtain ground truth (motion artifacts-free) volumes and corresponding displacement maps, pairs of horizontal and vertical 3D OCT volumes with motion artifacts are collected, and each volume is corrected with its orthogonal reference using the motion correction algorithm in [47], as illustrated in Fig. 4.3. Note that we use horizontal and vertical volume pairs for ground truth, but the proposed network takes only one single (horizontal or vertical) OCT volume as input.

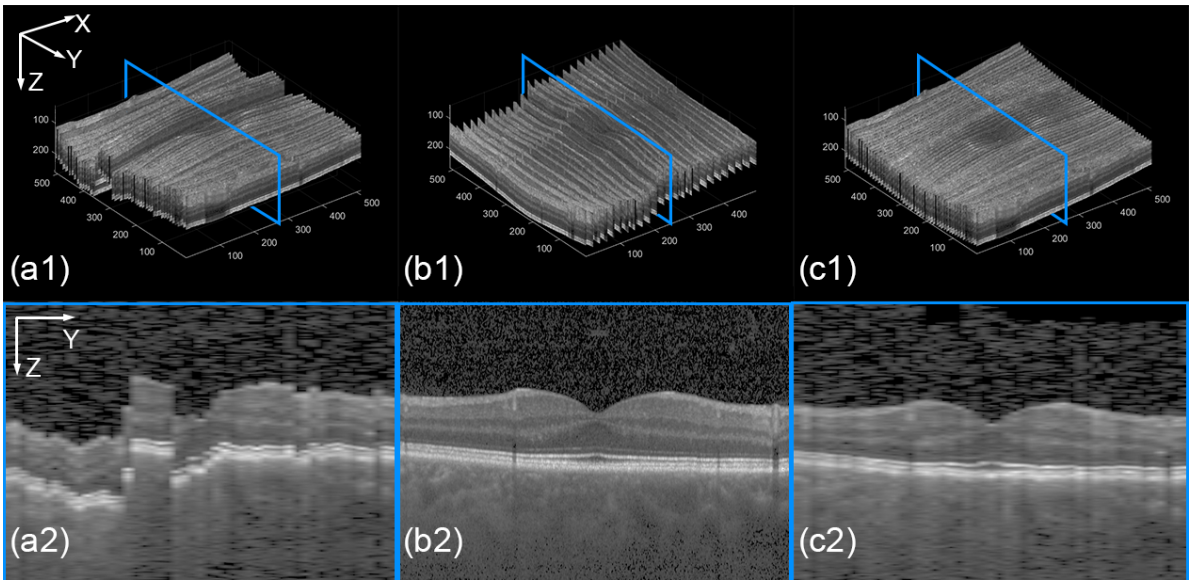


Figure 4.3. Orthogonal method [47] for ground truth acquisition. Column (a) shows the horizontal volume, (b) shows the paired vertical volume, (c) shows the motion-corrected horizontal volume using the motion correction algorithm in [47]. Row (1) shows the 3D volumes, and row (2) shows the cross-sectional B-scan.

4.3.3 Post processing with linear least squares

During the inference time, we apply a post-processing step where a linear function is fitted to the X axis of the predicted displacement $\mathbf{D}_z$ via linear least squares, in order to guarantee that the resulting fast B-scans in $\mathbf{V}_{\mathrm{dz}}$ have linear displacement in the Z direction. Specifically, denote the coordinates $\mathbf{X} \in \mathbb{R}^{W\times 2}$ as

$$\mathbf{X} = \begin{bmatrix} 0, & 1, & \cdots, & W-1 \\ 1, & 1, & \cdots, & 1 \end{bmatrix}^T, \tag{4.4}$$

the line parameters $\beta \in \mathbb{R}^{2\times N}$ can be obtained by solving the linear least squares problem

$$\beta^*(y) = \underset{\beta(y)\in\mathbb{R}^{2\times 1}}{\arg\min} \left\| \mathbf{D}_z(y) - \mathbf{X}\beta(y) \right\|^2. \tag{4.5}$$

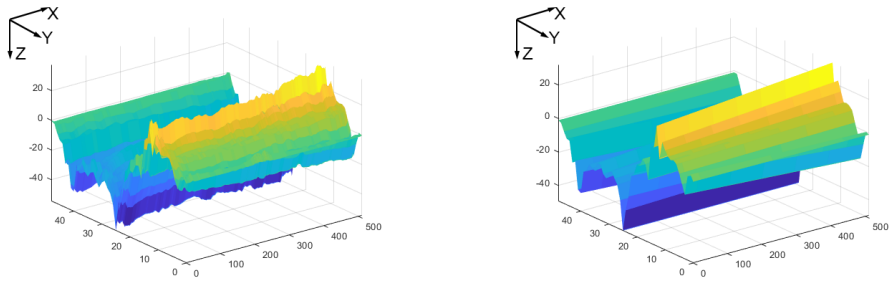
where $\beta(y)$ and $\mathbf{D}_z(y)$ denote the y-th row of β and $\mathbf{D}_z$, respectively. The solution is derived as

$$\beta(y) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}_z(y), \quad y = 0,\cdots,N-1 \tag{4.6}$$

and the displacement map obtained by solving the weighted least squares problem would be

$$\mathbf{D}'_z(y) = \mathbf{X}\beta(y) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}(y), \quad y = 0, \cdots, N-1. \tag{4.7}$$

Fig. 4.4 illustrates an example, where the displacement maps without post-processing and with post-processing are shown in (a) and (b), respectively. It can be observed that the noise in (a) is removed along the X axis after the least square line fitting step.



(a) Without post-processing **(b)** With post-processing

Figure 4.4. Post processing with linear least squares. (a) shows the displacement without post-processing, (b) shows the displacement after post-processing.

4.3.4 Loss function

The loss function consists of two loss terms, including a displacement L1 loss, and a displacement smoothness loss. Denoting the predicted displacement $\mathbf{D}_z$ and the ground truth displacement $\mathbf{D}_z^{\text{GT}} \in \mathbb{R}^{W\times N}$, the displacement L1 loss is given by

$$\mathscr{L}_{\text{disp}}(\mathbf{D}_z; \mathbf{D}_z^{\text{GT}}) = \text{mean}\Big(\mathbf{M} \odot |\mathbf{D}_z - \mathbf{D}_z^{\text{GT}}|\Big), \tag{4.8}$$

where $|\cdot|$ denotes the absolute value, $\odot$ is an element-wise multiplication, and $\mathbf{M} \in \mathbb{R}^{W\times N}$ is a predefined mask in [0,1], for assigning more weight at the center and less weight at the boundary of the OCT scan, as the center is the most important region of interest in clinical applications.

The second term is a displacement smoothness loss inspired by [128] to enforce smoothness

along the fast-scanning axis

$$\mathscr{L}_{\text{smooth}}(\mathbf{D}_z, \mathbf{D}'_z) = \text{mean}\left|\mathbf{D}_z - \mathbf{D}'_z\right|, \tag{4.9}$$

which is an L1 error between the raw displacement $\mathbf{D}_z$ and the least squares smoothed $\mathbf{D}'_z$ in eq. (4.7). Finally, the total loss is a weighted sum of the two loss terms

$$\mathscr{L} = \lambda_{\text{disp}}\mathscr{L}_{\text{disp}} + \lambda_{\text{smooth}}\mathscr{L}_{\text{smooth}}, \tag{4.10}$$

where λ_{disp} and λ_{smooth} are tunable weighting parameters.

4.3.5 Data augmentation

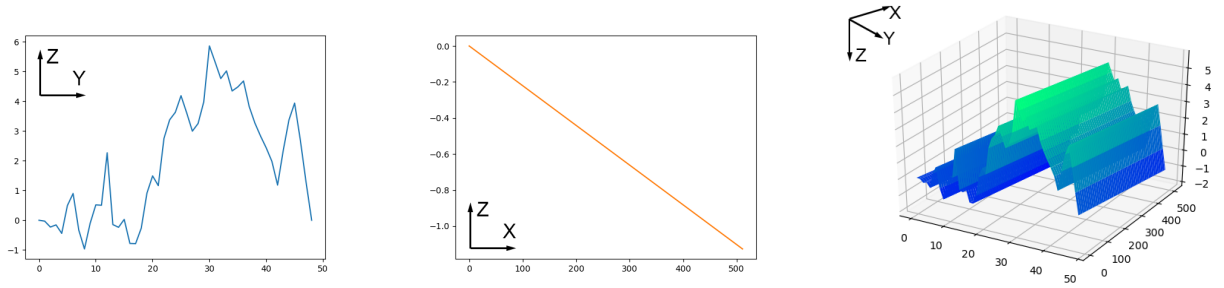
The input OCT volume and segmentation boundaries in the training data are augmented by random flipping on X and Y axes to prevent over-fitting. We also add random displacement for augmentation on top of the existing eye motion as follows: An N-dimensional Gaussian random vector with zero mean and unit variance $\mathbf{g} = (g_0, g_1, \cdots, g_N)^T$ is generated, and its cumulated sum is computed as

$$\left[g_0, \quad g_0 + g_1, \quad \cdots, \quad \textstyle\sum_{k=0}^{N-1} g_k\right]^T. \tag{4.11}$$

Then the tilt is removed from the axial augmentation on Y dimension $\delta_{\text{Y}} \in \mathbb{R}^{N\times 1}$ where $\delta_{\text{Y}}(0) = \delta_{\text{Y}}(N-1) = 0$, and the n-th element is

$$\delta_{\text{Y}}(n) = \sum_{k=1}^{n} g_k - \frac{n}{N-1}\sum_{k=1}^{N-1} g_k, \;\; n = 1, ..., N-2 \tag{4.12}$$

and an example of δ_{Y} is shown in Fig. 4.5(a). The axial augmentation on X dimension $\delta_{\text{X}} \in \mathbb{R}^{W\times 1}$ is generated by interpolating between 0 and a random number drawn from a Gaussian random variable with zero mean and unit variance, as shown Fig. 4.5(b). Finally, the total augmentation $\delta = \mathbf{1}_W \delta_{\text{Y}}^T + \delta_{\text{X}} \mathbf{1}_N^T$ is applied to the input OCT volume, where $\mathbf{1}_k \in \mathbb{R}^{k\times 1}$ denotes a k-dimensional vector of ones. Finally, δ_{Y} is subtracted from the ground truth displacement $\mathbf{D}_z^{\text{GT}} = \mathbf{D}_z^{\text{GT}} - \mathbf{1}_W \delta_{\text{Y}}^T$.



(a) Y dimension δ_{Y} **(b)** X dimension δ_{X} **(c)** Total δ

Figure 4.5. Data augmentation with random axial displacement. (a) Augmentation on Y dimension δ_{Y}, (b) augmentation on X dimension δ_{X}, (c) Total augmentation δ is applied to the input OCT volume.

4.4 Coronal motion correction network

In the proposed coronal motion correction network, we only focus on X motion for two reasons. Firstly, Y motion is very small compared to distance between neighboring B-scans based on statistics in [46]. Secondly, it is difficult to obtain ground truth of Y motion using conventional approaches, which will be discussed in subsection 4.4.1. The X motion correction network aims to predict a 1D displacement vector to the X axis $\mathbf{D}_x \in \mathbb{R}^{N\times 1}$, where N is the number of B-scans. Negative displacement shifts the B-scan left and positive displacement shifts it right. The magnitude of displacement denotes the number of pixel to be shifted and it is divided by a normalization factor X_{norm} for better numerical stability. Similar to the axial motion correction, the X motion corrected OCT volume $\mathbf{V}_{\mathrm{dx}}$ is derived by shifting the Z corrected volume $\mathbf{V}_{\mathrm{dz}}$ by $\mathbf{D}_x$ displacement to the X axis as

$$\mathbf{V}_{\mathrm{dx}}(z,x,y) = \mathbf{V}_{\mathrm{dz}}\left(z,\ x - \mathrm{int}\left(X_{\mathrm{norm}}\mathbf{D}_x(y)\right),\ y\right), \tag{4.13}$$

where $(x,y,z) \in [0,W-1]\times[0,N-1]\times[0,H-1]$ and X_{norm} is a normalization factor.

4.4.1 Ground truth acquisition

Many methods in the literature require registering and jointly optimizing multiple orthogonal or parallel OCT volumes [119, 53, 120, 40]. Furthermore, since their methods use a large number of B-scans in each OCT volume (e.g. $496\times512\times512$) to obtain C-scans with high-resolution in both fast and slow scanning axes, they fail to achieve visually desirable performance when resolution of

the slow axis is low, as in our dataset ($496 \times 512 \times 49$). Therefore, it is a major challenge to prepare the appropriate ground truth for training the coronal motion correction network.

The ground truth for X directional displacement $\mathbf{D}_x^{\mathrm{GT}}$ is obtained from the HEYEX software by Heidelberg Spectralis, which corrects residual X motion from the hardware eye-tracking system [21]. The magnitude of extracted X displacement is on average 0.96 pixel and 5 pixels at the maximum.

4.4.2 Network design

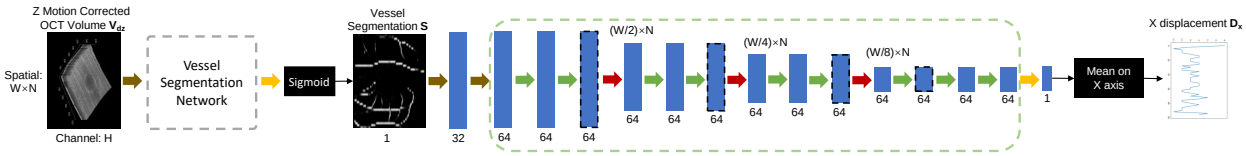


Figure 4.6. Network architecture of the vessel segmentation-based X motion correction network. The vessel segmentation network in dashed block adopts a similar structure as the dashed block in Fig. 4.2.

The network consists of a segmentation sub-network and a motion prediction sub-network. The segmentation sub-network first extracts a 2D vessel segmentation map on the en-face plane, where the coronal motion artifacts can be best observed. A subsequent motion prediction network then predicts a 1D displacement map $\mathbf{D}_x$ for motion to the X axis from the 2D vessel segmentation map.

Vessel segmentation

Since the X motion can be best observed by discontinuities of retinal vessel in OCT C-scans, and many related works also extract vessels for coronal motion correction [119, 40], we apply a vessel segmentation sub-network before the X motion prediction sub-network to extract critical information. Illustrated by gray dashed block in Fig. 4.6, the vessel segmentation network adopts a U-Net architecture similar to Fig. 4.2, which takes the Z motion corrected volume $\mathbf{V}_{\mathrm{dz}}$ to obtain the en-face C-scan vessel segmentation $\mathbf{S}$.

In order to train the OCT vessel segmentation sub-network, we utilize the IR multimodal information captured concurrently with the OCT volume, which is aligned with the OCT en-face

C-scan. In Fig. 4.7, we apply the IR vessel segmentation network [129] to extract the probability of vessel (vesselness) from the IR image, and it is downsampled to $W \times N$ to match the resolution of OCT C-scan. The downsampled segmentation of the IR image is then converted to binary label by thresholding at 0.5 and used as ground truth $\mathbf{S}^{\mathrm{GT}}$ for training the OCT vessel segmentation sub-network.

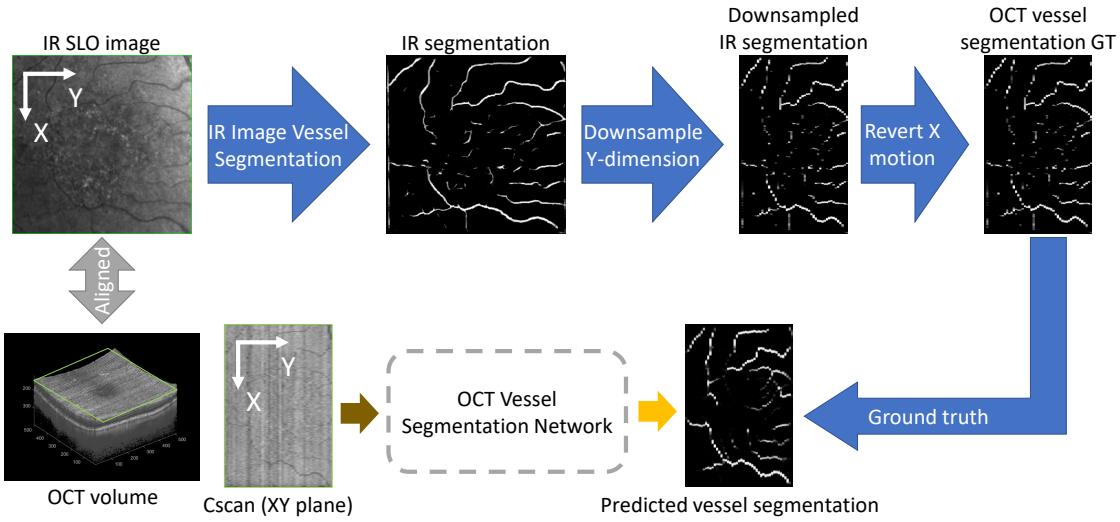


Figure 4.7. Procedure to obtain ground truth for training the OCT vessel segmentation sub-network.

The segmentation sub-network is first trained with a combination of binary cross-entropy loss and soft Dice loss. The binary cross-entropy loss can be expressed as

$$\begin{aligned}\mathscr{L}_{\mathrm{BCE}}(\mathbf{S},\mathbf{S}^{\mathrm{GT}}) = -\mathrm{mean}\Big(&\mathbf{S}^{\mathrm{GT}}\log\mathbf{S} \\ &+(1-\mathbf{S}^{\mathrm{GT}})\log(1-\mathbf{S})\Big).\end{aligned} \tag{4.14}$$

Note that the Sigmoid function for the predicted segmentation $\mathbf{S}$ can be integrated into the binary cross-entropy loss using the "log-sum-exp" trick for better numeric stability. The soft Dice loss is defined as one minus the soft Dice coefficient

$$\mathscr{L}_{\mathrm{Dice}}(\mathbf{S},\mathbf{S}^{\mathrm{GT}}) = 1 - \frac{2\sum\mathbf{S}\odot\mathbf{S}^{\mathrm{GT}}}{\sum\mathbf{S}+\sum\mathbf{S}^{\mathrm{GT}}}, \tag{4.15}$$

where $\odot$ denotes elementwise product, which measures the overlapping ratio between predicted

vessel and ground truth. Finally, the vessel segmentation sub-network is trained with a total loss and $\lambda_{\text{BCE}} = \lambda_{\text{Dice}} = 1$,

$$\mathscr{L}_{\text{vessel}} = \lambda_{\text{BCE}}\mathscr{L}_{\text{BCE}} + \lambda_{\text{Dice}}\mathscr{L}_{\text{Dice}}. \tag{4.16}$$

X motion prediction

After training the vessel segmentation sub-network, we freeze the weights in the vessel segmentation network and train a X motion prediction sub-network to predict X displacement based on the 2D vessel segmentation map. The input to the X motion prediction sub-network is $1 \times W \times N$ with B batch size. In the X motion prediction network, the resolution on the fast scanning axis is reduced by 3 downsampling convolutions with kernel size 2×1 and stride 2×1. The network output is synthesized by averaging across the fast scanning axis to obtain $\mathbf{D}_x$. Since the X displacement is a 1D vector whose element is a single displacement for each B-scan slice, we simplify the upsampling branch in the U-Net structure. The MSE Loss function is applied between predicted and ground truth displacement with normalization factor X_{norm}:

$$\mathscr{L}_{\text{xdisp}} = \underset{y}{\text{mean}}||X_{\text{norm}}\mathbf{D}_x(y) - \mathbf{D}_x^{\text{GT}}(y)||^2. \tag{4.17}$$

4.5 Experimental result

In the experiment, we compare the axial and coronal motion correction performance of our proposed network to that of four other methods [48, 54, 123, 21] which operate on a single OCT volume input.

4.5.1 Dataset

We evaluate the performance of motion correction algorithms on two datasets collected by Jacobs Retina Center. The first dataset consists of 99 eyes with paired horizontal and vertical OCT volumes which are obtained by Heidelberg Spectralis in an imaging volume of $1.9 \times 5.8 \times 5.8$ (mm^3) with 20 degree field of view. Hardware eye-tracking is always turned on during the imaging process. All the OCT scans come with instrument's segmentation boundaries of 11 retinal layers.

Among 99 horizontal and 99 vertical volumes, the dimensions of 9 volumes are $496 \times 512 \times 25$, while those of the remaining 189 volumes are $496 \times 512 \times 49$. The 198 OCT volumes (99 horizontal and 99 vertical) are divided into 142, 19, 37 for training, validation, and testing, respectively. The dataset include both healthy subjects as well as patients with wet and dry AMD, diabetic retinopathy, and other diseases including epi-retinal membrane, macular edema, retinal detachment, macular hole, chorioretinopathy, and posterior vitreous detachment.

The second dataset includes only 106 singular OCT volumes (horizontal direction only). This dataset is also captured with the Heidelberg Spectralis OCT system, but covers a wider range of resolutions: there are 95 OCT volumes of the same resolution $496 \times 512 \times 49$ in the training dataset, while there are also 3 volumes of resolution $496 \times 512 \times 25$, 6 volumes of resolution $496 \times 1024 \times 97$, and 2 volumes of resolution $496 \times 1024 \times 49$.

4.5.2 Tilt correction

We apply a tilt correction algorithm after axial motion correction for all methods at inference time, which leads to better visualization of OCT volume as illustrated in Fig. 4.8. The motion corrected OCT volumes before and after tilt correction are shown in Fig. 4.8 row (1) and (2), respectively. The four corners of the RPE segmentation boundary in sub-image (c1) are mapped to the same reference plane at height $H_{\text{ref}} = 300$ by applying another displacement $\mathbf{D}_z^{\text{tilt}}$,

$$\begin{aligned}\mathbf{D}_z^{\text{tilt}}(x,y) = H_{\text{ref}} - \frac{1}{(W-1)(N-1)}\Big(&\mathbf{B}(1,0,0)(W-1-x)(N-1-y)\\ &+\mathbf{B}(1,W-1,0)(N-1-y)x+\mathbf{B}(1,0,N-1)(W-1-x)y\\ &+\mathbf{B}(1,W-1,N-1)xy\Big),\end{aligned}$$

which is based on bilinear interpolation of the four corner coordinates. Finally, the tilt corrected ILM and RPE surfaces are visualized in sub-image (c2).

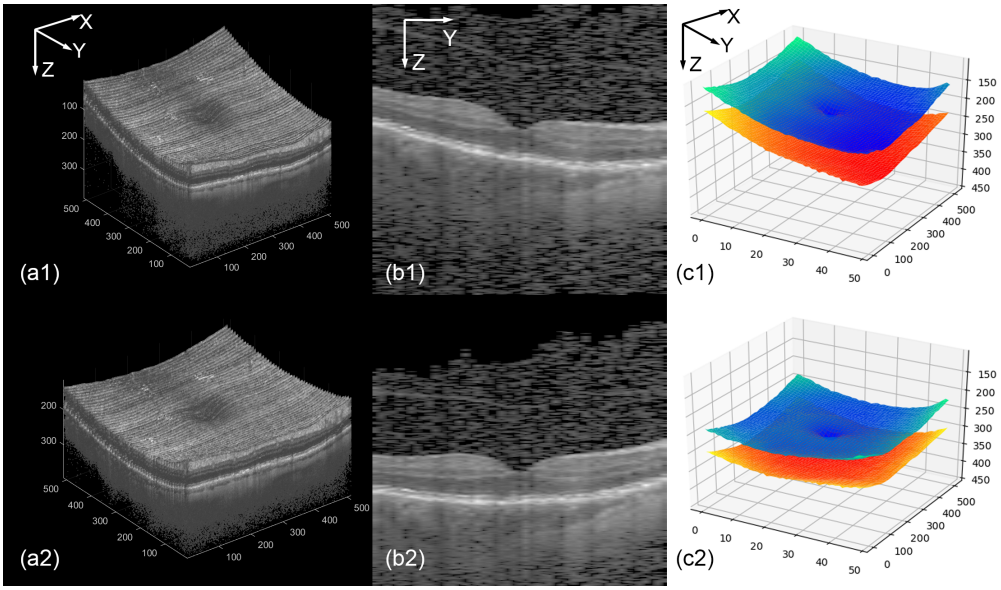


Figure 4.8. Tilt correction on motion corrected OCT volume. Row (1) shows the original OCT and row (2) shows the tilt corrected result. The 3D OCT volume, the cross-sectional B-scan, and ILM and RPE surfaces are presented in the columns (a), (b), and (c), respectively.

4.5.3 Criteria

We evaluate the algorithms based on two criteria: the pixel-wise mean absolute error (MAE) and the smoothness of OCT scan. The first metric is MAE between the predicted and ground truth displacement, which measures the overall accuracy and preservation of retinal curvature, where smaller MAE indicates better performance.

$$\mathrm{MAE}_z(\mathbf{D}_z;\mathbf{D}_z^{\mathrm{GT}}) = Z_{\mathrm{norm}}\underset{x,y}{\mathrm{mean}}\Big(|\mathbf{D}_z(x,y)-\mathbf{D}_z^{\mathrm{GT}}(x,y)|\Big). \tag{4.18}$$

Similarly, the MAE for X motion is

$$\mathrm{MAE}_x(\mathbf{D}_x;\mathbf{D}_x^{\mathrm{GT}}) = X_{\mathrm{norm}}\underset{y}{\mathrm{mean}}\Big(|\mathbf{D}_x(y)-\mathbf{D}_x^{\mathrm{GT}}(y)|\Big). \tag{4.19}$$

We also include an intensity based metric, Pearson correlation coefficient (PCC) [107] to measure the similarity between motion corrected OCT volume $\mathbf{V}^{\mathrm{pred}}$ (could be $\mathbf{V}_{\mathrm{dx}}$ or $\mathbf{V}_{\mathrm{dz}}$ depending on the experiment) and ground truth OCT volume $\mathbf{V}^{\mathrm{GT}}$. The PCC, which is a commonly used metric

in literature [46], is defined as

$$\mathrm{PCC}(\mathbf{V}^{\mathrm{pred}}, \mathbf{V}^{\mathrm{GT}}) = \frac{\mathrm{cov}(\mathbf{V}^{\mathrm{pred}}, \mathbf{V}^{\mathrm{GT}})}{\sigma_{\mathbf{V}^{\mathrm{pred}}} \sigma_{\mathbf{V}^{\mathrm{GT}}}}, \tag{4.20}$$

where $\sigma_{\mathbf{V}^{\mathrm{pred}}}$ and $\sigma_{\mathbf{V}^{\mathrm{GT}}}$ respectively are the standard deviations of volume $\mathbf{V}^{\mathrm{pred}}$ and $\mathbf{V}^{\mathrm{GT}}$, and the covariance can be calculated by

$$\mathrm{cov}(\mathbf{V}^{\mathrm{pred}}, \mathbf{V}^{\mathrm{GT}}) = \mathrm{E}[(\mathbf{V}^{\mathrm{pred}} - \mathrm{E}[\mathbf{V}^{\mathrm{pred}}])(\mathbf{V}^{\mathrm{GT}} - \mathrm{E}[\mathbf{V}^{\mathrm{GT}}])]. \tag{4.21}$$

In order to evaluate the preservation of retinal curvature, we adopt the curvature index proposed by [122]. The central B-scan (X-Z plane at $Y = \mathrm{floor}(N/2)$) and the central cross sectional B-scan (Y-Z plane at $X = \mathrm{floor}(W/2)$) are taken to evaluate the X and Y curvatures, respectively. For each direction, the tilt is first corrected, and the RPE curvature is obtained by fitting a 4-th order polynomial to the segmentation boundary of the RPE using least squares. For the X direction, the polynomial coefficients $\mathbf{p}_x$ are

$$\mathbf{p}_x = \underset{\mathbf{p}\in\mathbb{R}^{5\times 1}}{\arg\min} \sum_{x=0}^{W-1} \left| \mathbf{B}\left(1, x, \mathrm{floor}\left(\frac{N}{2}\right)\right) - [1,\ x,\ x^2,\ x^3,\ x^4]\mathbf{p} \right|^2, \tag{4.22}$$

and similarly for the Y direction,

$$\mathbf{p}_y = \underset{\mathbf{p}\in\mathbb{R}^{5\times 1}}{\arg\min} \sum_{y=0}^{N-1} \left| \mathbf{B}\left(1, \mathrm{floor}\left(\frac{W}{2}\right), y\right) - [1,\ y,\ y^2,\ y^3,\ y^4]\mathbf{p} \right|^2. \tag{4.23}$$

Then the curvature index Curv is defined as the ratio between length of the RPE curve and the distance of a straight line after converting pixels to millimeters [122]. Denoting the OCT resolution in millimeters with L_z, L_x, L_y, (in our case $1.9 \times 5.8 \times 5.8$ (mm^3) as mentioned in subsection A.)

$$\mathrm{Curv}_x = \frac{L_z \cdot \mathrm{length}([1,\ x,\ x^2,\ x^3,\ x^4]\mathbf{p}_x)}{H \cdot L_x}, \tag{4.24}$$

$$\text{Curv}_y = \frac{L_z \cdot \text{length}([1,\ y,\ y^2,\ y^3,\ y^4]\mathbf{p}_y)}{H \cdot L_y}. \tag{4.25}$$

Then the distortion of curvature Dist to each direction is defined as the L1 difference between the curvature index of ground truth and predicted OCT volumes,

$$\text{Dist}_t = |\text{Curv}_t^{\text{pred}} - \text{Curv}_t^{\text{GT}}|,\ \ t \in \{x, y\}. \tag{4.26}$$

As observed in [122], the X and Y directional curvature should be relatively similar. When ground truth Y curvature is not available, we evaluate L1 error between Y curvature Curv_y and the ground truth (same as the input) X curvature $\text{Curv}_x^{\text{GT}}$

$$\text{Dist}_{xy} = |\text{Curv}_y^{\text{pred}} - \text{Curv}_x^{\text{GT}}|. \tag{4.27}$$

Finally, we use the Dice coefficient between the corrected and ground truth vessel segmentation map in the C-scan to evaluate the performance of X motion correction. Denoting the binary segmentation maps as $\mathbf{S}_1$ and $\mathbf{S}_2$, the Dice coefficient can be obtained by

$$\text{Dice}(\mathbf{S}_1, \mathbf{S}_2) = \frac{2 \times \sum(\mathbf{S}_1 \odot \mathbf{S}_2)}{\sum \mathbf{S}_1 + \sum \mathbf{S}_2}. \tag{4.28}$$

The Dice coefficient is a value between [0,1], and higher value shows higher overlapping ratio with ground truth.

4.5.4 Implementation

The proposed networks are implemented in PyTorch. For Z motion correction, both the baseline network and the network with segmentation input have 484K model parameters. Reflected padding is used for convolutions and dropouts with $p = 0.2$ is applied on every resolution level. We set $Z_{\text{norm}} = 10$ for the normalization factor. The models are trained with Adam optimizer with weight decay 10^{-3} and batch size 4. An initial learning rate of 10^{-3} and exponential decay with

momentum 0.995 are set to train the network for 500 max epochs and $\lambda_{\text{disp}} = 1$, $\lambda_{\text{smooth}} = 0.5$ for loss function. The best model, which is selected based on the lowest validation loss, is applied for the test set.

For the X motion correction, the vessel segmentation sub-network is first trained for 500 max epochs, using Adam optimizer with weight decay 10^{-3}, batch size 4, initial learning rate 10^{-3}, and exponential decay with momentum 0.99. The best model selected on the validation set achieves an average segmentation accuracy of 96.05% and Dice coefficient of 0.4776. After freezing the model parameters in the vessel segmentation network, the X motion prediction sub-network is trained for 1000 max epochs, using Adam optimizer with weight decay 10^{-3}, batch size 4, and learning rate 10^{-4}. We set $X_{\text{norm}} = W/512$ for the normalization factor depending on the B-scan resolution, and use $p = 0.4$ for dropouts. Standard data augmentation techniques including random cropping and flipping are included during training. Finally, the best model is selected on the validation set. At inference time, we convert the predicted displacement vector to integer pixels to avoid interpolation within fast B-scans.

We compare the proposed method with several other methods, which only take a single OCT volume to correct motion. Since the method in [47] requires orthogonal OCT volume pairs to obtain ground truth, it is excluded in the comparison. Two comparison methods [48, 54] are implemented in Python. The method in [123] is implemented in MATLAB based on the original authors' implementation of saliency detection [130].

4.5.5 Evaluation on Test Dataset

In the first experiment, we evaluate the axial and coronal motion correction methods on the testset with 37 volumes. The average inference time to correct one OCT volume is 0.29 seconds on a NVIDIA GeForce RTX 2080 Ti GPU. The qualitative results for axial motion correction using different methods are shown in Fig. 4.9. Rows (1-3) show the experimental results of one example OCT volume with moderate motion, while rows (4-6) show another example with larger motion. Rows (1,4) show the 3D volumes and rows (2,5) show the cross-sectional B-scans on the Y-Z plane

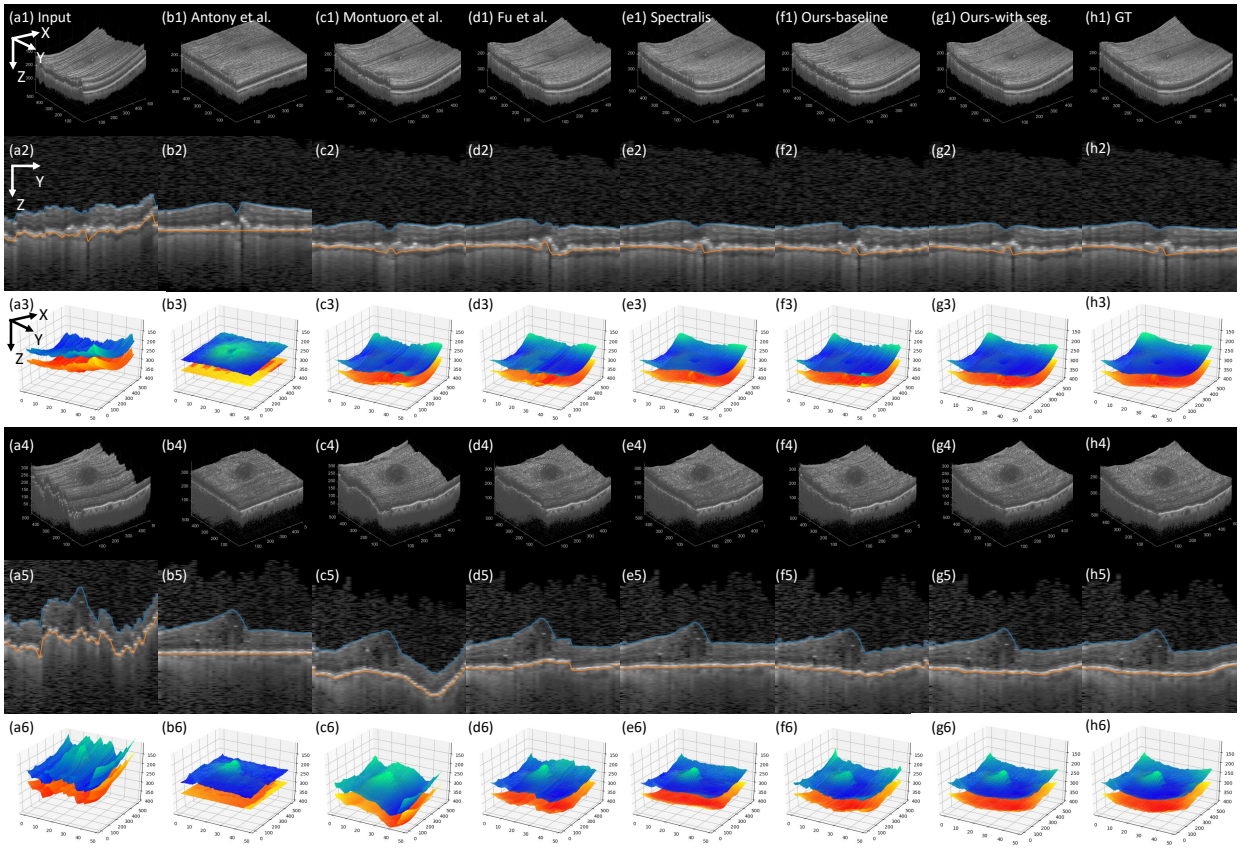


Figure 4.9. Qualitative result of different Z motion correction methods on the test set. Row (1-3) and (4-6) show two examples. Row (1,4) show the 3D volume, row (2,5) show the cross-sectional B-scans with segmentation boundaries of RPE and ILM, and row (3,6) show the segmentation boundaries of RPE and ILM in 3D.

with segmentation boundaries of RPE and ILM where they are overlaid onto the image in blue and orange lines, respectively, and Gamma correction at 2.2 is applied for visualization. Rows (3,6) show the segmentation boundaries of RPE and ILM after correction in 3D.

The input OCT volumes are visualized in column (a), and column (b) shows the method by Antony et al. [54], which flattens the RPE and results in errors when the disease alters RPE in (b2). The method proposed by Montuoro et al. [48] of column (c) is able to correct axial motion without flattening the retina, but it results in unnatural curvature at large motion in (c2) and (c4) that is not similar to the ground truth in column (h). The results of Fu et al. [123] in column (d) are smooth in most B-scans, but the errors lead to abrupt discontinuities. The software correction of Spectralis system [21] in column (e) can effectively smooth the motion, whereas the recovered

curvature is different from the ground truth obtained from orthogonal OCT pairs. Finally, our proposed baseline network in column (f) can reduce the motion artifacts in the input volume with some residual motion, while the network with segmentation input in column (g) yields smoother correction result while recovering the overall curvature.

Table 4.1. Quantitative result of different axial (Z) motion correction on the test set.

Method		Z correction only						
		MAE_z	$PCC(\mathbf{V}_{dz}, \mathbf{V}_{GT})$	$Curv_x$	$Dist_x$	$Curv_y$	$Dist_y$	$Dist_{xy}$
Before correction		22.742 (20.703)	0.5505 (0.149)	1.0020 (0.002)	-	-	-	-
Ground truth		-	1.0000 (0.000)	1.0020 (0.002)	-	1.0013 (0.002)	-	0.0010 (0.001)
Antony et al. [54]		40.530 (17.118)	0.3517 (0.114)	1.0000 (0.000)	0.0020 (0.002)	1.0000 (0.000)	0.0012 (0.002)	0.0020 (0.002)
Montuoro et al. [48]		20.543 (19.853)	0.5666 (0.152)	**1.0020** (0.002)	**0.0000** (0.000)	1.0051 (0.008)	0.0040 (0.007)	0.0044 (0.007)
Fu et al. [123]		20.963 (18.394)	0.5605 (0.140)	**1.0020** (0.002)	**0.0000** (0.000)	1.0048 (0.007)	0.0043 (0.007)	0.0042 (0.006)
Spectralis [21]		15.253 (9.746)	0.6000 (0.140)	**1.0020** (0.002)	**0.0000** (0.000)	1.0003 (0.001)	0.0010 (0.001)	0.0017 (0.002)
Ours (baseline)	-	13.948 (10.833)	0.6185 (0.112)	1.0017 (0.002)	0.0007 (0.001)	**1.0013** (0.002)	0.0009 (0.001)	0.0013 (0.001)
	LS	12.720 (9.504)	0.6388 (0.114)	**1.0020** (0.002)	**0.0000** (0.000)	1.0016 (0.002)	0.0009 (0.001)	**0.0012** (0.001)
Ours (with seg.)	-	9.102 (7.972)	0.7150 (0.097)	1.0012 (0.001)	0.0009 (0.002)	1.0009 (0.001)	0.0008 (0.001)	0.0015 (0.002)
	LS	**8.512** (8.211)	**0.7277** (0.111)	**1.0020** (0.002)	**0.0000** (0.000)	1.0008 (0.001)	**0.0006** (0.001)	0.0014 (0.002)

The quantitative results of different Z motion correction are shown in Table 4.1, where each entry shows the mean and standard deviation value. We first evaluate the input OCT volumes before correction as baseline, and we also compute the curvature for ground truth. As the method by Antony et al. [54] flattens the retina, it achieves a curvature index of 1.0000 for both X and Y direction, and it is the only method that distorts the curvature of the X direction ($Dist_x = 0.0020$). It also yields larger MAE than the input (no correction) by removing the retinal curvature. The method of Montuoro et al. [48] and Fu et al. [123] reduce the MAE and improve PCC compared with the input by a small margin, but their Y curvature are still very different from the ground truth ($Dist_y = 0.0040$, $Dist_y = 0.0043$ respectively). The software correction result of Spectralis [21] produces a MAE of 15.253 and PCC of 0.6000, and the Y curvature is flatter compared with ground truth. Overall, our method with segmentation and least squares (denoted by LS) post-processing achieves the lowest MAE at 8.512 pixels and the highest PCC at 0.7277, and the Y distortion $Dist_y$ is the smallest. Our baseline network without segmentation input achieves a higher MAE and lower PCC compared with the proposed network with segmentation, ranking as the second method. The average Y curvature of our baseline network is most similar to ground truth, but the average

distortion is larger compared with the network with segmentation input.

We also include an ablation study of the post-processing step, where we evaluate the raw network output displacement and the post-processed result with least squares (LS) in Table 4.1. The results show that LS post-processing lowers the MAE of both the baseline network from 13.948 to 12.720, and the network with segmentation input from 9.102 to 8.512. The PCC of the two networks increase compared to the ground truth, demonstrating that LS post-processing also improves the smoothness of corrected volume.

Table 4.2. Evaluation of different X motion correction networks on the test set.

	GT Z correction + X correction only		
Method	MAE_x	PCC	Dice
No correction	0.7484 (0.449)	0.9714 (0.019)	0.9205 (0.051)
Ground truth	-	1.0000 (0.000)	1.0000 (0.000)
Antony et al. [54]	0.7484 (0.449)	0.9714 (0.019)	0.9205 (0.051)
Montuoro et al. [48]	1.6152 (0.460)	0.9506 (0.015)	0.8191 (0.056)
Fu et al. [123]	0.8185 (0.438)	0.9693 (0.019)	0.9127 (0.049)
Spectralis [21]	-	-	-
Our X-network	**0.7406** (0.437)	**0.9716** (0.019)	**0.9218** (0.048)

The qualitative results of X motion correction using different methods are shown in Fig. 4.10. The en-face X-Y view of the two examples (same as Fig. 4.9) are shown in rows (1-2) and (2-4). The en-face projection C-scans shown in rows (1,3) are obtained by averaging from the ELM to RPE layer [37], while the vessel segmentation is the intermediate output of our proposed X motion correction network. Since the corrected result is shown in cyan and the ground truth is assigned in the red channel, the overlapping region appears in gray. We also note the Dice coefficient of the overlaid segmentation maps in Fig. 4.10. The IR SLO images and their vessel segmentation maps [129] are presented in column (g) as reference for the true shape of vessels. Since the method proposed by Antony et al. [54] only includes Z motion correction, it is excluded in this comparison. The methods by Montuoro et al. [48] and Fu et al. [123] fail to reduce the x motion such that the Dice values of the methods are even lower than one of input (No correction). Our proposed method with both the baseline network and the network with segmentation are able to increase the Dice coefficient and reduce the error.

The quantitative results of X motion correction is shown in Table 4.2. In order to evaluate X

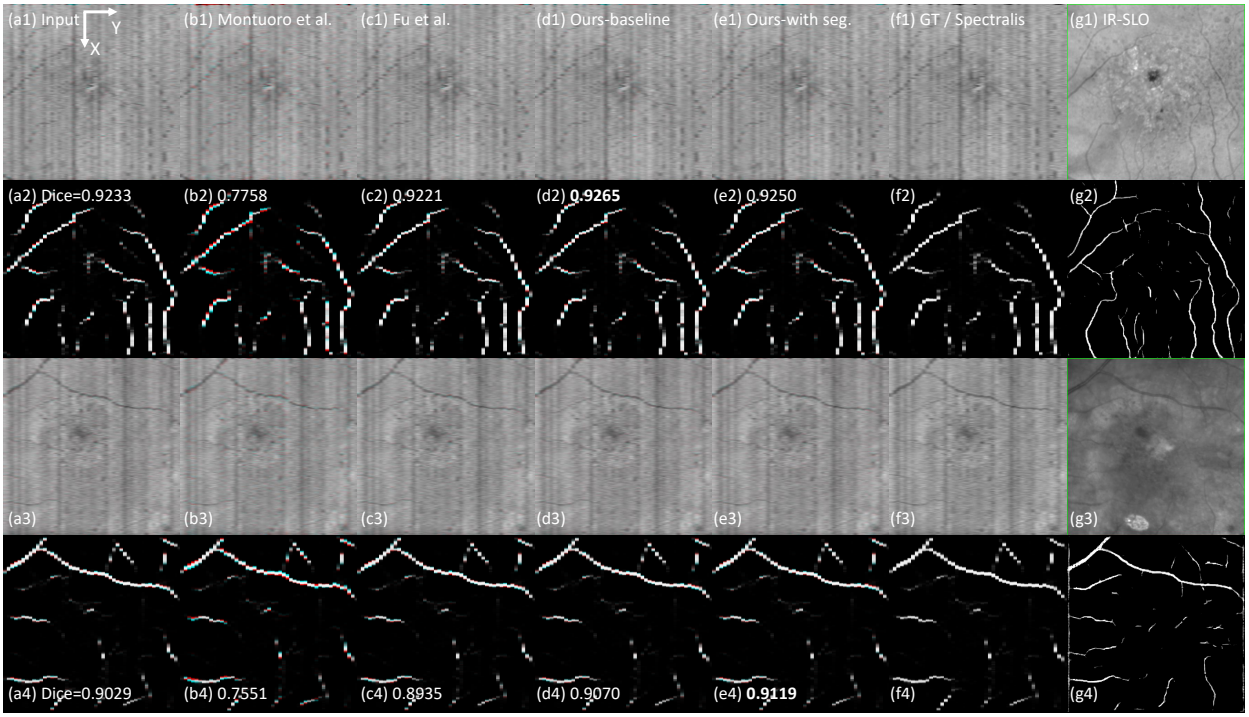


Figure 4.10. Qualitative result of different X motion correction methods on the test set. Row (1,3) show the overlay of C-scan of corrected OCT volumes and ground truth, and row (2,4) show the segmentation overlay. The ground truth is shown in the red channel, and the corrected result is shown in cyan. Reference IR SLO images and their segmentation maps are shown in column (g). Please zoom in to see details.

motion correction performance alone without the influence of different Z motion correction methods, we include an ablation study using ground truth Z correction input for the X motion correction step in each method. The amplitude of X motion is very small compared with Z motion in this ablation study, however, it can still be observed that only our method reduces the X motion in the input volume, achieving the lowest MAE and highest Dice coefficient. The results also demonstrate that the methods of Montuoro et al. [48] and Fu et al. [123] fail motion correction, reflected by higher MAE and lower Dice coefficient compared with the input (No correction).

Finally, we analyze the performance of joint Z and X motion correction algorithms based on diseases in Table 4.3. We first test the performance with joint Z and X motion correction for each method, denoted by "All data" in Table 4.3. The test set is classified into the following categories: wet and dry Age-related Macular Degeneration (AMD), Diabetes, other diseases (foveal contour distortion), and normal. Overall, our proposed methods achieve the best motion correction

performance for all diseases, shown by lower MAE in both Z and X direction, higher PCC, higher Dice coefficient compared with ground truth. Our proposed methods also achieve the best curvature preservation for each type of disease, demonstrated by Curv_x and Curv_y closer to the ground truth and smaller Dist_x, Dist_y, and Dist_{xy} values.

Table 4.3. Quantitative result of different motion correction methods on the test set with different diseases.

				Joint Z + X correction					
	Metrics	Before correction	GT	Antony et al. [54]	Montuoro et al. [48]	Fu et al. [123]	Spectralis [21]	Ours (baseline)	Ours (with seg.)
All data	MAE_z	22.742 (20.703)	-	40.530 (17.118)	20.543 (19.853)	20.962 (18.394)	15.253 (9.746)	12.720 (9.504)	**8.5118** (8.211)
	MAE_x	0.7484 (0.449)	-	-	1.8507 (0.544)	0.8232 (0.414)	-	**0.7428** (0.440)	0.7401 (0.440)
	PCC	0.5505 (0.149)	-	0.3517 (0.114)	0.5635 (0.151)	0.5604 (0.140)	0.6043 (0.141)	0.6388 (0.114)	**0.7277** (0.111)
	Dice	0.9205 (0.051)	-	-	0.7663 (0.070)	0.9119 (0.046)	-	**0.9214** (0.049)	0.9219 (0.049)
	Curv_x	1.0020 (0.002)	1.0020 (0.002)	1.0000 (0.000)	**1.0020** (0.002)	**1.0020** (0.002)	**1.0020** (0.002)	**1.0020** (0.002)	**1.0020** (0.002)
	Dist_x	-	-	0.0020 (0.002)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)
	Curv_y	-	1.0013 (0.002)	1.0000 (0.000)	1.0051 (0.008)	1.0048 (0.007)	1.0003 (0.001)	**1.0016** (0.002)	1.0008 (0.001)
	Dist_y	-	-	0.0012 (0.002)	0.0040 (0.007)	0.0043 (0.007)	0.0010 (0.001)	0.0009 (0.001)	**0.0006** (0.001)
	Dist_{xy}	-	-	0.0020 (0.002)	0.0044 (0.007)	0.0042 (0.006)	0.0017 (0.002)	**0.0012** (0.001)	0.0014 (0.002)
Wet AMD	MAE_z	34.120 (22.993)	-	48.060 (25.124)	30.506 (21.708)	27.466 (11.263)	23.886 (11.221)	20.040 (12.736)	**15.288** (12.868)
	MAE_x	1.0796 (0.243)	-	-	2.0673 (0.297)	1.1265 (0.227)	-	1.0735 (0.236)	**1.0673** (0.240)
	PCC	0.4866 (0.112)	-	0.3673 (0.191)	0.5080 (0.131)	0.5249 (0.087)	0.5338 (0.112)	0.5894 (0.115)	**0.6705** (0.147)
	Dice	0.8780 (0.035)	-	-	0.7389 (0.054)	0.8733 (0.032)	-	0.8792 (0.033)	**0.8803** (0.034)
	Curv_x	1.0039 (0.003)	1.0039 (0.003)	1.0000 (0.000)	**1.0039** (0.003)	**1.0039** (0.003)	**1.0039** (0.003)	**1.0039** (0.003)	**1.0039** (0.003)
	Dist_x	-	-	0.0039 (0.003)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)
	Curv_y	-	1.0027 (0.002)	1.0000 (0.000)	1.0086 (0.013)	1.0088 (0.006)	1.0008 (0.001)	**1.0031** (0.002)	1.0013 (0.001)
	Dist_y	-	-	0.0027 (0.002)	0.0064 (0.011)	0.0072 (0.005)	0.0019 (0.002)	0.0016 (0.002)	**0.0015** (0.002)
	Dist_{xy}	-	-	0.0039 (0.003)	0.0070 (0.010)	0.0061 (0.005)	0.0031 (0.003)	**0.0014** (0.002)	0.0026 (0.003)
Dry AMD	MAE_z	11.804 (6.017)	-	45.505 (3.690)	11.121 (6.369)	12.051 (4.964)	15.767 (6.265)	10.019 (3.987)	**5.5237** (2.252)
	MAE_x	0.9218 (0.146)	-	-	2.1088 (0.340)	0.9626 (0.116)	-	**0.9150** (0.124)	**0.9150** (0.120)
	PCC	0.6354 (0.087)	-	0.3502 (0.010)	0.6313 (0.101)	0.6230 (0.071)	0.5718 (0.106)	0.6425 (0.101)	**0.7663** (0.069)
	Dice	0.9073 (0.027)	-	-	0.7440 (0.070)	0.9026 (0.026)	-	0.9082 (0.023)	**0.9085** (0.022)
	Curv_x	1.0015 (0.001)	1.0015 (0.001)	1.0000 (0.000)	**1.0015** (0.001)	**1.0015** (0.001)	**1.0015** (0.001)	**1.0015** (0.001)	**1.0015** (0.001)
	Dist_x	-	-	0.0015 (0.001)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)
	Curv_y	-	1.0008 (0.000)	1.0000 (0.000)	1.0029 (0.002)	1.0012 (0.002)	1.0000 (0.000)	1.0005 (0.001)	**1.0007** (0.000)
	Dist_y	-	-	0.0008 (0.000)	0.0021 (0.002)	0.0013 (0.002)	0.0008 (0.000)	**0.0003** (0.000)	**0.0003** (0.000)
	Dist_{xy}	-	-	0.0015 (0.001)	0.0025 (0.002)	0.0015 (0.001)	0.0014 (0.001)	0.0012 (0.001)	**0.0011** (0.001)
Diabetes	MAE_z	46.258 (29.401)	-	36.490 (5.383)	42.440 (31.524)	34.825 (37.760)	9.7601 (2.963)	15.162 (9.132)	**7.6517** (4.250)
	MAE_x	0.9694 (0.140)	-	-	2.3214 (0.390)	1.0867 (0.091)	-	**0.9541** (0.128)	0.9643 (0.136)
	PCC	0.3495 (0.139)	-	0.3025 (0.039)	0.3687 (0.148)	0.4743 (0.204)	0.6248 (0.081)	0.5715 (0.109)	**0.6948** (0.114)
	Dice	0.9028 (0.019)	-	-	0.7489 (0.070)	0.8909 (0.016)	-	**0.9060** (0.018)	0.9039 (0.018)
	Curv_x	1.0007 (0.001)	1.0007 (0.001)	1.0000 (0.000)	**1.0007** (0.001)	**1.0007** (0.001)	**1.0007** (0.001)	**1.0007** (0.001)	**1.0007** (0.001)
	Dist_x	-	-	0.0007 (0.001)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)
	Curv_y	-	1.0005 (0.000)	1.0000 (0.000)	1.0079 (0.007)	1.0091 (0.013)	1.0000 (0.000)	1.0023 (0.002)	**1.0006** (0.000)
	Dist_y	-	-	0.0005 (0.000)	0.0075 (0.007)	0.0090 (0.012)	0.0004 (0.000)	0.0018 (0.001)	**0.0003** (0.000)
	Dist_{xy}	-	-	0.0007 (0.001)	0.0073 (0.006)	0.0086 (0.012)	0.0006 (0.001)	0.0019 (0.001)	**0.0004** (0.000)
Others	MAE_z	23.744 (3.940)	-	37.313 (11.924)	19.286 (3.597)	19.624 (4.495)	12.659 (6.965)	11.626 (6.179)	**5.8620** (2.543)
	MAE_x	0.9898 (0.181)	-	-	1.8316 (0.416)	1.0306 (0.185)	-	**0.9796** (0.184)	**0.9796** (0.184)
	PCC	0.4997 (0.029)	-	0.3674 (0.013)	0.5304 (0.037)	0.5481 (0.054)	0.6416 (0.133)	0.6499 (0.082)	**0.7660** (0.061)
	Dice	0.8985 (0.008)	-	-	0.7788 (0.066)	0.8939 (0.010)	-	**0.8997** (0.007)	**0.8997** (0.007)
	Curv_x	1.0010 (0.001)	1.0010 (0.001)	1.0000 (0.000)	**1.0010** (0.001)	**1.0010** (0.001)	**1.0010** (0.001)	**1.0010** (0.001)	**1.0010** (0.001)
	Dist_x	-	-	0.0010 (0.001)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)
	Curv_y	-	1.0008 (0.001)	1.0000 (0.000)	1.0082 (0.007)	1.0043 (0.003)	1.0000 (0.000)	1.0014 (0.000)	**1.0008** (0.000)
	Dist_y	-	-	0.0008 (0.001)	0.0073 (0.007)	0.0035 (0.003)	0.0008 (0.001)	0.0007 (0.000)	**0.0005** (0.000)
	Dist_{xy}	-	-	0.0010 (0.001)	0.0071 (0.006)	0.0034 (0.003)	0.0010 (0.001)	**0.0009** (0.000)	**0.0009** (0.000)
Normal	MAE_z	11.494 (8.603)	-	34.674 (13.856)	10.878 (8.487)	16.219 (16.053)	10.864 (7.048)	7.9207 (4.214)	**5.7582** (2.287)
	MAE_x	0.2712 (0.378)	-	-	1.4260 (0.554)	0.3807 (0.344)	-	0.2711 (0.364)	**0.2649** (0.360)
	PCC	0.6379 (0.129)	-	0.3507 (0.087)	0.6451 (0.129)	0.5891 (0.168)	0.6556 (0.164)	0.6925 (0.100)	**0.7524** (0.080)
	Dice	0.9715 (0.039)	-	-	0.7990 (0.066)	0.9580 (0.037)	-	0.9715 (0.038)	**0.9725** (0.037)
	Curv_x	1.0014 (0.001)	1.0014 (0.001)	1.0000 (0.000)	**1.0014** (0.001)	**1.0014** (0.001)	**1.0014** (0.001)	**1.0014** (0.001)	**1.0014** (0.001)
	Dist_x	-	-	0.0014 (0.001)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)
	Curv_y	-	1.0007 (0.001)	1.0000 (0.000)	1.0015 (0.002)	1.0021 (0.006)	1.0001 (0.000)	**1.0006** (0.001)	**1.0006** (0.000)
	Dist_y	-	-	0.0007 (0.001)	0.0009 (0.002)	0.0023 (0.006)	0.0007 (0.001)	0.0005 (0.001)	**0.0003** (0.000)
	Dist_{xy}	-	-	0.0014 (0.001)	0.0015 (0.002)	0.0029 (0.006)	0.0013 (0.001)	**0.0009** (0.001)	0.0010 (0.001)

4.5.6 Evaluation on Dataset with Various Resolutions

In the second experiment, we test the performance of the Z and X motion correction methods on another dataset with only horizontal scan OCT volumes and different resolutions. For X motion correction, we still utilize the Spectralis software correction result as ground truth. The MAE for X displacement and Dice coefficient between ground truth and corrected segmentation maps are evaluated. Since we cannot obtain axial correction ground truth via orthogonal registration from the single volume dataset, we only evaluate the distortion of curvature to the X and Y direction. Due to lack of ground truth, we evaluate the curvature distortion to the Y direction based on Dist_{xy} in eq. (4.27). Note that we derive the curvature distortion to the X direction Dist_x in the same way as the previous experiment using eq. (4.26).

In Table 4.4, we compare the evaluation results of joint Z and X motion correction on the whole dataset of 106 images, as well as on the subset of each resolution. For evaluation of X motion correction result on the whole dataset, our proposed methods can reduce the MAE and increase the Dice coefficient compared with the input without any correction, and the baseline network yields a slightly better result. The methods proposed by Montuoro et al. [48] and Fu et al. [123] on the other hand increase the MAE and decrease the Dice coefficient compared with the input, which introduce larger X motion artifacts. When analyzing for each resolution, our proposed method still achieves the best X motion correction performance for most resolutions, except on the resolution of $496 \times 1024 \times 49$. Fu et al. [123] achieves the best performance for this resolution, while our methods remain the same as input. This may be due to the lack of data, as there are only 2 examples of this resolution. For Z motion correction, the numeric range of curvature index is consistent with our previous dataset. While the average Y curvature of our baseline network is more similar to the X curvature reference for all resolutions except the resolution of $496 \times 512 \times 25$, our network with segmentation input achieves smaller average Y distortion for all resolutions except $496 \times 1024 \times 49$. Overall, our proposed methods achieve the best preservation of retinal curvature.

Table 4.4. Quantitative result of different motion correction methods on dataset with different resolutions.

		Joint Z + X correction						
Category	Metrics	Before correction	Antony et al. [54]	Montuoro et al. [48]	Fu et al. [123]	Spectralis [21]	Ours (baseline)	Ours (with seg.)
All data	MAE_x	0.8630 (0.227)	-	8.1138 (6.219)	0.9156 (0.226)	-	**0.8607** (0.226)	0.8619 (0.225)
	Dice	0.9200 (0.028)	-	0.6125 (0.132)	0.9158 (0.028)	-	**0.9204** (0.028)	0.9202 (0.028)
	$Curv_x$	1.0012 (0.001)	1.0000 (0.000)	**1.0012** (0.001)	**1.0012** (0.001)	**1.0012** (0.001)	**1.0012** (0.001)	**1.0012** (0.001)
	$Dist_x$	-	0.0012 (0.001)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)
	$Curv_y$	-	1.0000 (0.000)	1.0027 (0.004)	1.0027 (0.005)	1.0003 (0.001)	**1.0013** (0.001)	1.0007 (0.001)
	$Dist_{xy}$	-	0.0012 (0.001)	0.0025 (0.004)	0.0028 (0.004)	0.0012 (0.002)	0.0012 (0.001)	**0.0009** (0.001)
$496 \times 512 \times 49$	MAE_x	0.8997 (0.175)	-	7.6411 (5.243)	0.9527 (0.180)	-	**0.8988** (0.172)	0.8990 (0.172)
	Dice	0.9155 (0.024)	-	0.6169 (0.132)	0.9111 (0.024)	-	**0.9157** (0.024)	0.9156 (0.024)
	$Curv_x$	1.0012 (0.001)	1.0000 (0.000)	**1.0012** (0.001)	**1.0012** (0.001)	**1.0012** (0.001)	**1.0012** (0.001)	**1.0012** (0.001)
	$Dist_x$	-	0.0012 (0.001)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)
	$Curv_y$	-	1.0000 (0.000)	1.0024 (0.003)	1.0025 (0.004)	1.0003 (0.001)	**1.0012** (0.001)	1.0007 (0.001)
	$Dist_{xy}$	-	0.0012 (0.001)	0.0022 (0.003)	0.0025 (0.004)	0.0012 (0.002)	0.0012 (0.002)	**0.0009** (0.001)
$496 \times 512 \times 25$	MAE_x	0.0000 (0.000)	-	10.350 (4.089)	0.1067 (0.038)	-	**0.0000** (0.000)	**0.0000** (0.000)
	Dice	1.0000 (0.000)	-	0.5449 (0.078)	0.9938 (0.004)	-	1.0000 (0.000)	1.0000 (0.000)
	$Curv_x$	1.0013 (0.000)	1.0000 (0.000)	**1.0013** (0.000)	**1.0013** (0.000)	**1.0013** (0.000)	**1.0013** (0.000)	**1.0013** (0.000)
	$Dist_x$	-	0.0013 (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)
	$Curv_y$	-	1.0000 (0.000)	1.0158 (0.011)	1.0039 (0.004)	1.0016 (0.002)	1.0027 (0.001)	**1.0015** (0.001)
	$Dist_{xy}$	-	0.0013 (0.000)	0.0145 (0.011)	0.0031 (0.004)	0.0023 (0.001)	0.0014 (0.001)	**0.0007** (0.000)
$496 \times 1024 \times 97$	MAE_x	0.7663 (0.097)	-	15.238 (13.442)	0.8058 (0.092)	-	**0.7388** (0.110)	0.7577 (0.106)
	Dice	0.9378 (0.008)	-	0.5308 (0.115)	0.9346 (0.009)	-	**0.9414** (0.004)	0.9390 (0.006)
	$Curv_x$	1.0017 (0.002)	1.0000 (0.000)	**1.0017** (0.002)	**1.0017** (0.002)	**1.0017** (0.002)	**1.0017** (0.002)	**1.0017** (0.002)
	$Dist_x$	-	0.0017 (0.002)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)
	$Curv_y$	-	1.0000 (0.000)	1.0014 (0.002)	1.0073 (0.010)	1.0001 (0.000)	**1.0011** (0.001)	1.0003 (0.000)
	$Dist_{xy}$	-	0.0017 (0.002)	0.0016 (0.002)	0.0071 (0.009)	0.0016 (0.002)	**0.0014** (0.001)	**0.0014** (0.002)
$496 \times 1024 \times 49$	MAE_x	0.7041 (0.092)	-	5.8414 (1.168)	**0.6939** (0.102)	-	0.7041 (0.092)	0.7041 (0.092)
	Dice	0.9626 (0.002)	-	0.7477 (0.056)	**0.9629** (0.002)	-	0.9626 (0.002)	0.9626 (0.002)
	$Curv_x$	1.0020 (0.002)	1.0000 (0.000)	**1.0020** (0.002)	**1.0020** (0.002)	**1.0020** (0.002)	**1.0020** (0.002)	**1.0020** (0.002)
	$Dist_x$	-	0.0020 (0.002)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)	**0.0000** (0.000)
	$Curv_y$	-	1.0000 (0.000)	1.0002 (0.000)	1.0007 (0.000)	1.0003 (0.000)	**1.0014** (0.001)	1.0008 (0.001)
	$Dist_{xy}$	-	0.0020 (0.002)	0.0018 (0.002)	0.0018 (0.001)	0.0017 (0.002)	**0.0009** (0.001)	0.0012 (0.001)

4.6 Conclusion

In this paper, we proposed end-to-end deep learning neural networks for axial and coronal OCT motion correction using only a single OCT volume. The proposed baseline axial motion correction network adopts a modified U-Net architecture to predict a Z displacement map, which can be enhanced with the OCT layer segmentation. We also proposed a coronal motion network, which includes a vessel segmentation sub-network and an X motion prediction sub-network. The experimental results show that the proposed method is able to correct axial and coronal motion while recovering the retinal curvature and achieving significant improvement compared to the conventional methods even with disease or large motion artifacts at various resolutions.

In future work, we can generalize the proposed method to different OCT systems and investigate the possibility of training with simulation data to reduce dependency on large dataset with ground truth. The proposed method will be integrated to better display and visualization of 3D OCT scans and benefit subsequent processing including retinal layer segmentation and OCT-A imaging.

Chapter 4, in part, is a reprint of the material as it appears in IEEE International Conference on Image Processing 2021 (ICIP), Y. Wang, A. Warter, M. Cavichini, W. R. Freeman, D. G. Bartsch, T. Q. Nguyen, C. An, IEEE, 2021. The dissertation author is the primary author of this paper.

Chapter 4, in part, has been submitted for publication of the material as it may appear in IEEE Transactions on Image Processing, Y. Wang, A. Warter, M. Cavichini, V. Alex, D. G. Bartsch, W. R. Freeman, T. Q. Nguyen, C. An, IEEE, 2021. The dissertation author is the primary author of this paper.

Chapter 5

OCT layer segmentation

5.1 Introduction

Optical Coherence Tomography (OCT) is a 3D imaging technology widely used in ophthalmology. An infrared beam is used to obtain the cross-sections of the retina in vivo at high resolution [112]. The role of OCT imaging is crucial in diagnosing and monitoring both retinal and systemic diseases [2], including age-related macular degeneration (AMD), diabetic macular edema (DME), glaucoma, multiple sclerosis (MS), and so on.

In OCT imaging, the back-scattered intensities of infrared beam represent 1D depth (A-scan, Z axis of Fig. 5.1). By moving the beam in a raster scanning pattern, a sequence of 2D cross-sectional images (B-scan, XZ plane of Fig. 5.1) can be acquired. Finally, the 3D OCT volume can be formed by stacking the B-scans (XZ planes) to the Y axis. The *fast scanning axis* refers to the direction where a B-scan is acquired (X axis of Fig. 5.1), and the *slow scanning axis* refers to the direction where B-scans are stacked (Y axis of Fig. 5.1).

Cross-sectional imaging of OCT is useful for observing the layered structures of the retina, and changes of the retinal layers are critical indicators of both retinal and systemic diseases [131]. For example, thinning of the retinal nerve fiber layer (RNFL) and ganglion cell layer (GCL) is frequently used for assessment of glaucoma. The overall retinal thickness is often used for assessment of DME and choroidal neovascularization (CNV) [2]. It is therefore important to develop an accurate segmentation for retinal layers to assess these changes automatically. In particular,

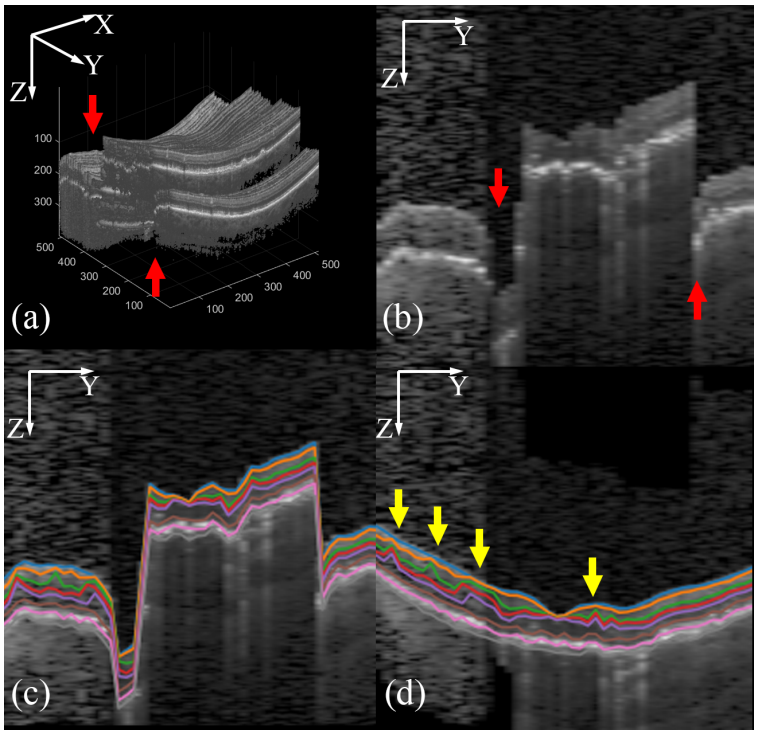


Figure 5.1. OCT motion artifacts and segmentation. (a) The axial motion artifacts in 3D OCT volume indicated with red arrows, (b) slow B-scan (YZ plane) with motion artifacts, (c) 2D segmented layers with OCT motion artifacts, (d) 2D segmented layers after OCT motion correction, with 3D inconsistency indicated by yellow arrows.

recent studies reveal that the thickness and vessel density of RNFL is related to Alzheimer disease and Parkinson's disease [1, 15], and joint OCT-A vessel density estimation with layer segmentation algorithm could be used to develop clear and and non-invasive tool for early detection of these CNS disorders.

Many OCT layer segmentation approaches have been proposed [58, 59, 132, 50], and commercial OCT systems also provided their own segmentation softwares [21]. However, segmentation error is prevalent with these approaches and compromises the quality of downstream tasks such as OCT-A projections [49]. Recent deep learning segmentation neural networks led to significant improvement of accuracy with publicly available annotated datasets [60, 4, 62, 64, 133, 65]. Most networks are modified based on the 2D U-net [66] architecture, which has achieved remarkable performance in numerous image segmentation tasks. However, most deep learning methods applied on the 2D B-scan slides often ignore the 3D contextual information within neighboring B-scans.

Motion artifact is one of the major reasons that hinder the development of 3D contextual information [46]. Motion artifacts in OCT can be caused by involuntary head motion, respiration, pulsation, or fixational eye movements during the imaging process [46]. These involuntary motions lead to axial and coronal misalignment between neighboring B-scans, shown in sub-figures (b) and (c) in Fig. 5.1, respectively, where major motion artifacts are indicated by red arrows. The axial motion introduces discontinuities in the cross-sectional B-scan as in sub-figure (b), which results in discontinuities in the 2D segmented layers in sub-figure (c). After motion correction, the discontinuities are reduced in sub-figure (d), but the layers lack 3D consistency.

The first segmentation approach utilizing 3D information was proposed by Garvin et al. (OCTExplorer) [67, 134], which applies 3D "feasibility" constraints to reduce failures of 2D graph-based approach. Nevertheless, the motion artifacts are removed by flattening the bottom surface of the retina, which also removes the retinal curvature. Besides constraining the 2D segmentation, 3D information can also be used for denoising upon correction of motion artifacts. In the RETOUCH OCT Fluid Detection and Segmentation challenge [135], the winner team [136] performs bounded variation 3D smoothing to reduce speckle noise as pre-processing. However, the 3D information is not fully utilized as their segmentation network is still trained on 2D slices.

In this chapter, we propose a deep learning method that combines motion correction and 3D segmentation. A motion correction neural network first corrects the axial motion artifacts in the input OCT volume, and then a graph-assisted 3D neural network is trained with 3D input and 3D output. The performance of the proposed method is compared with commercial OCT software solutions, as well as several state-of-the-art methods in literature. Experimental results show that motion correction and 3D contextual information enhance the accuracy of the OCT layer segmentation.

5.2 Related work

With the development of OCT imaging systems in the past two decades, many OCT segmentation methods have been proposed. However, segmentation errors in cases of various diseases

and motion artifacts are still prevalent with existing segmentation algorithms [49].

Most existing OCT layer segmentation algorithms are 2D image-based, meaning that the segmentation predictions are based on a single B-scan, and applied slice-by-slice for 3D data [56, 132, 50, 58, 59, 137]. Some methods predict the 1D boundaries between each retinal layer, while other methods predict the 2D pixel-wise label. The advantages of predicting 1D boundaries include topology guaranteed layers (i.e. the first boundary will always be above the second boundary) and robustness to outlier regions. However, the 1D boundary method is not able to precisely characterize the layers in retinal disease such as the example illustrated in Fig. 5.2. Even for human corrected segmentation boundaries, it is not possible to precisely follow the shape of the disease to the one-to-one mapping nature of the 1D boundaries.

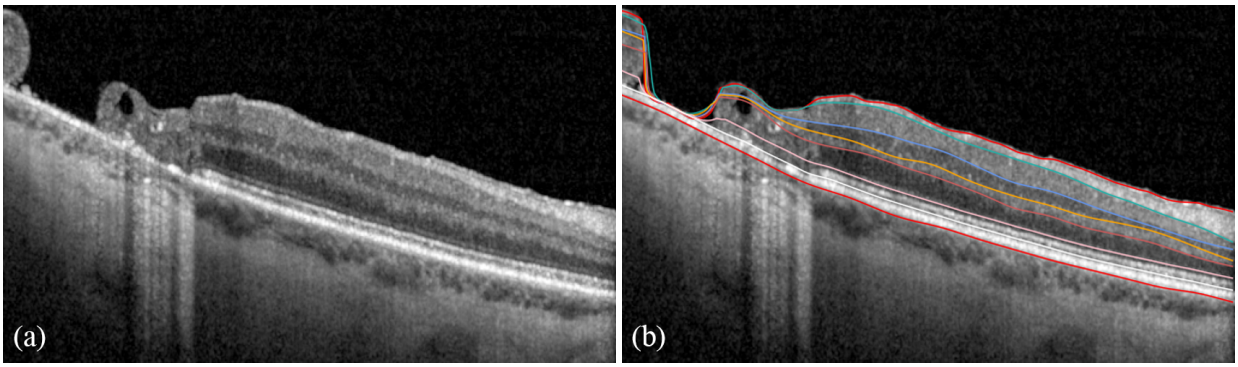


Figure 5.2. Limitations of 1D segmentation boundaries. (a) An example OCT B-scan with retinal disease, (b) human corrected 1D segmentation boundaries. The boundaries could not precisely follow the shape of the disease.

Conventional methods that predict the 1D boundaries include level set methods [58, 59], but these methods take extremely long computational time for up to hours per OCT volume. Graph-based methods are another popular category of algorithm [132, 138, 137], which post-process the pixel-wise prediction from machine learning classifiers [50, 60]. However, these conventional methods are difficult to generalize to various retinal diseases, and require manually established features and extensive parameter tuning.

Thanks to several public datasets with available annotation [50, 36, 139], deep learning-based methods have improved the accuracy of the pixel-wise label prediction via end-to-end training [60, 4, 62, 64, 133, 65]. RelayNet [140] is one of the first deep learning application in retinal

layer segmentation. It modifies the 2D U-Net architecture, and uses the weighted cross-entropy loss to penalize error near each boundary. Other methods combine deep learning classifiers with conventional post-processing to obtain layer boundaries from pixel-wise prediction. Fang et al. [60] combined a patch-based neural network with graph search post-processing to segment 9 layer boundaries. Pekala et al. [65] proposed a dense U-net classifier with post-processing using Gaussian process regression to segment 5 layer boundaries. He et al. [62] proposed to use a second neural network to correct topology based on initial U-Net prediction. Some methods aim to predict the 1D boundary based on end-to-end regression networks. He et al. [64] proposed cascaded U-Nets to learn the thickness map of each layer achieving topology constraints. The architecture was later improved [63, 133] by multi-task training and including X, Y coordinates as input. The latest state-of-the-art method is the MGU-Net [139] which combines U-Net with graph-convolution inspired global reasoning blocks [141], achieving the highest Dice coefficient reported on the DME dataset [50].

A major problem that impedes the development of 3D segmentation approaches is that involuntary motion causes misalignment artifacts between neighboring B-scans in 3D OCT imaging. Therefore, motion correction is required to recover the motion-free 3D OCT volume. Some OCT systems integrate eye-tracking hardware to compensate for eye-motion, and there are also post-processing algorithms to correct motion after OCT acquisition [46]. Axial movement is observed to be more significant than coronal movement in magnitude [47], and the higher axial resolution also result in larger axial shift in pixels [47, 53]. Hence, many methods solely focus on correction of axial motion between B-scans [134, 54, 125].

The first OCT segmentation approach to utilize 3D information was proposed by Garvin et al. (OCTExplorer) [67, 134], which could segment 7 boundaries for macular centered OCT scans. The approach first flattens the bottom surface of the retina to remove motion artifacts (along with retinal curvature) and then enhances 2D graph-based methods by additional 3D "feasibility" constraints. The feasibility constraints take advantage of 3D contextual information and enforce smoothness in neighboring surfaces and surface distance constraints. They demonstrate that their proposed

method with 3D information could reduce segmentation failure compared to the 2D graph-based approaches. Besides conventional approaches, DeepMind [4] proposed a 3D segmentation network taking 9 consecutive B-scans that could segment 15 classes of features to aid disease classification. The major limitation is that only two retinal layers can be identified by the segmentation network, namely the neurosensory retina and the RPE. It is therefore promising to combine motion correction and 3D neural networks in retinal OCT segmentation.

In this chapter, we proposed a 3D OCT segmentation pipeline that combines two neural networks to correct motion artifacts and perform segmentation based on volumetric data, which enables utilization of 3D contextual information to achieve improved accuracy. To the best of our knowledge, this is one of the first 3D deep learning methods applied in the retinal layer segmentation task.

5.3 Proposed method

In this chapter, we propose to combine OCT motion correction network with a graph-assisted 3D neural network for retinal layer segmentation. The proposed 3D segmentation pipeline is illustrated in Fig. 5.3. In the motion correction stage, a 2D segmentation method is first applied to the input OCT volume $\mathbf{V}$ slice-by-slice to obtain the binary segmentation $\mathbf{S}^{\text{bin}}$ for retinal and non-retinal regions. The motion correction network [125] then takes the 3D volume $\mathbf{V}$ and the extracted top (inner limiting membrane, ILM) and bottom (Bruch's membrane, BM) layer boundary as input, to predict a 2D displacement map $\mathbf{D}$ that compensates for axial motion. The original 3D volume $\mathbf{V}$ is warped using the 2D displacement map $\mathbf{D}$ to obtain the motion-corrected volume $\mathbf{V}'$. In the second stage, the specific retinal layers $\mathbf{S}^{\text{pred}}$ are classified using a segmentation network based on the 3D motion-corrected OCT volume $\mathbf{V}'$.

The axial motion correction network [125] takes the raw OCT volume $\mathbf{V} \in \mathbb{R}^{H \times W \times N}$ as input, where H, W, and N denotes the resolution along the Z, X, and Y axes. The BM and ILM boundaries are also included for improved performance. The output of the network is a 2D displacement map $\mathbf{D} \in \mathbb{R}^{W \times N}$, which compensates for Z directional motion in the 3D OCT volume. The architecture

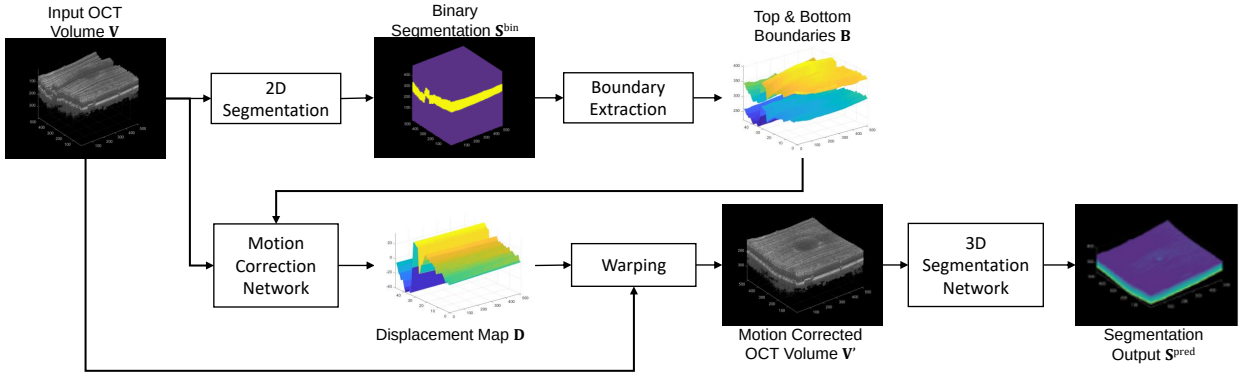


Figure 5.3. Proposed 3D OCT segmentation pipeline with motion correction.

of the network is modified based on a residual U-Net, and the performance is verified for various diseases and resolutions [125]. After network prediction, the motion-corrected OCT volume $\mathbf{V}'$ can be obtained by warping the input volume $\mathbf{V}$ based on the predicted axial displacement map $\mathbf{D}$.

$$\mathbf{V}'(z,x,y) = \mathbf{V}(z - \mathbf{D}(x,y),x,y). \tag{5.1}$$

After correcting the motion artifacts, the proposed network takes M neighboring B-scans in the motion-corrected OCT volume as input and predicts the segmentation in K classes for the M B-scans as shown in Fig. 5.4. The segmentation network is inspired by the U-net [66] and MGU-Net [139], which includes analysis and synthesis of 4 resolutions. The 3D convolution blocks are separated into spatial $3 \times 3 \times 1$ convolutions and depth-wise $1 \times 1 \times 3$ convolutions, which utilizes 3D information without significantly increasing the number of parameters. Specifically, a graph pyramid structure is used at the lowest resolution for global reasoning. The graph pyramid is comprised of 4 resolutions with graph reasoning units (GRU) [141]. Each graph reasoning units include three branches for projection to node space, re-projection to feature space, and fusion of global features, where the two graph convolution blocks are used after projection to the node space. The standard convolution on image data with grid coordinates could be interpreted as a nearest neighbor graph. However, by utilizing the top projection branch of GRU, the input features could be projected to latent space using a learned projection matrix, which enables global reasoning over disjoint and distant areas. After projection to the latent node space, a graph is obtained where each

node contains a feature state. Two graph convolutions are performed implemented by channel-wise and node-wise 1D convolutions. Finally, the graph is re-projected from node space to image space using a learned inverse projection matrix.

For comparison of the effect of 3D convolution, we train a 2D version of the segmentation network by removing the $1 \times 1 \times 3$ convolutions and replacing all the 3D operations with their 2D counterparts. We also use the 2D segmentation network to derive the top and bottom layer boundaries for the motion correction network.

The proposed 3D architecture includes 1.940M trainable parameters, and the 2D version includes 1.909M parameters. Both are reduced compared with MGU-Net which has 2.094M parameters, yet the graph pyramid structure can effectively improve the segmentation performance as demonstrated in the experimental result.

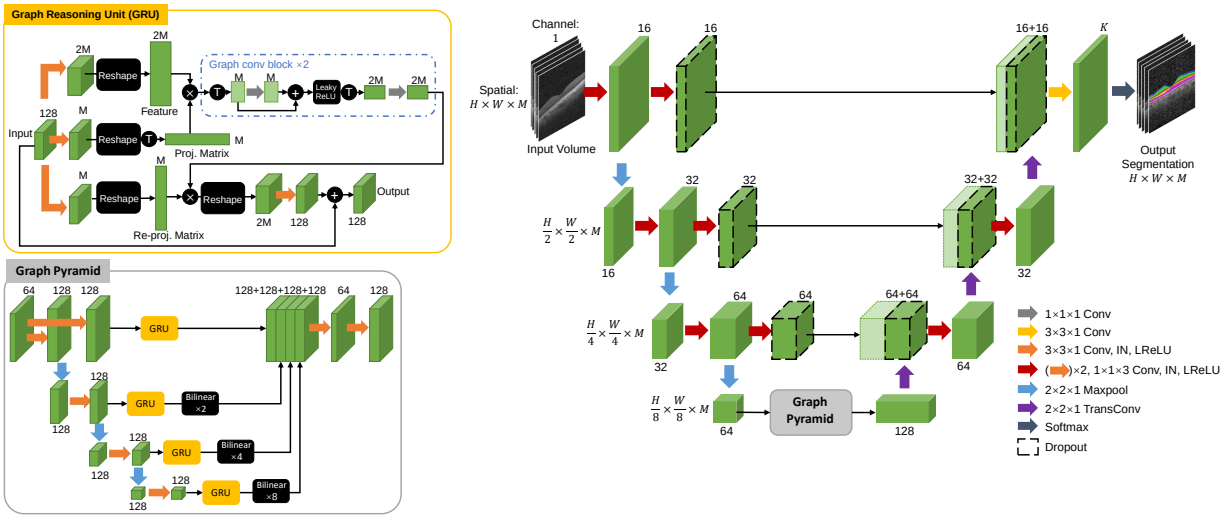


Figure 5.4. Proposed 3D OCT segmentation network with graph pyramid architecture. Here "IN" operation denotes Instance Normalization, "LReLU" denotes LeakyReLU activation, "T" in black circle denotes transpose, and "×" in black circle denotes matrix multiplication.

The segmentation network is trained using a hybrid loss function, which is a weighted sum of cross-entropy loss and Dice loss. Denoting the last convolution layer output as $\mathbf{x} \in \mathbb{R}^{H \times W \times M}$, the ground truth class label as $\mathbf{y} \in \mathbb{R}^{H \times W \times M}$, and the ground truth one-hot label as $\mathbf{S}^{\mathrm{GT}}$. Note that we also include a valid mask $\mathbf{M} \in \mathbb{R}^{H \times W \times M}$ to exclude regions without annotation, where 1 denotes

annotated pixels and 0 denotes otherwise. The masked cross-entropy loss can be expressed as

$$\mathscr{L}_{\text{CE}}(\mathbf{x},\mathbf{y}) = \frac{-1}{\sum_n \mathbf{M}_n} \sum_n \Big(\log \frac{\exp \mathbf{x}_{\mathbf{y}_n,n}}{\sum_{k=0}^{K-1} \exp \mathbf{x}_{k,n}} \cdot \mathbf{M}_n \Big), \tag{5.2}$$

where n spans the batch and spatial dimensions. Note that the cross-entropy loss could be implemented with better numeric stability by using the "log-sum-exp" trick, combining log-softmax activation with the negative log likelihood loss for the last convolution layer output $\mathbf{x}$.

The Dice loss is included to regularize the class-imbalance issue of each retinal layer, and emphasize retinal region (class $k = 1$ to $K-2$) over non-retinal regions ($k = 0$ or $K-1$). It is defined as one minus the soft Dice coefficient, which is a score between 0 and 1 characterizing the overlapping ratio between prediction and ground truth. The soft Dice coefficient can be expressed as

$$\text{SoftDice}_k(\mathbf{x},\mathbf{S}^{\text{GT}}) = \frac{2\sum_n \sigma(\mathbf{x}_{k,n})\mathbf{S}^{\text{GT}}_{k,n}\mathbf{M}_n}{\sum_n \sigma(\mathbf{x}_{k,n})\mathbf{M}_n + \sum_n \mathbf{S}^{\text{GT}}_{k,n}\mathbf{M}_n}, \tag{5.3}$$

where $\sigma(\cdot)$ denotes the softmax function along the channel dimension. Then the Dice loss for retinal layers is defined as

$$\mathscr{L}_{\text{Dice}}(\mathbf{x},\mathbf{S}^{\text{GT}}) = 1 - \operatorname*{mean}_{k=1:K-2} \text{SoftDice}_k(\mathbf{x},\mathbf{S}^{\text{GT}}). \tag{5.4}$$

Finally, the total loss is combined with weights $\lambda_{\text{CE}} = 1$, $\lambda_{\text{Dice}} = 2$,

$$\mathscr{L}_{\text{total}} = \lambda_{\text{CE}}\mathscr{L}_{\text{CE}} + \lambda_{\text{Dice}}\mathscr{L}_{\text{Dice}}. \tag{5.5}$$

Simulated shearing along the X axis is used besides standard data augmentation methods such as random cropping and horizontal flipping, which adds another degree of freedom to the image transformation. The method to generate simulated shearing is described in [125]. The random shearing is generated by an affine transformation with two Gaussian variables $a \sim N(0, \sqrt{2/W})$

and $b \sim N(0,1)$,

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ a & 0 & 1 & b \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \tag{5.6}$$

Boundary extraction is a optional post-processing to convert the pixel-wise label prediction $\mathbf{S}^{\text{pred}}$ into segmentation boundaries $\mathbf{B}^{\text{pred}}$. The boundaries are detected, joined, and interpolated using the pseudo-code in Algorithm 1.

Algorithm 1. Pixel-wise label to boundary

Pixel-wise label $\mathbf{S}^{\text{pred}} \in \mathbb{R}^{H \times W}$
K+1 boundaries $\mathbf{B}^{\text{pred}} \in \mathbb{R}^{(K+1) \times W}$
for $k = 0 : K$ **do**
  k-th edge $\mathbf{C}$ (W lists)
  for $x = 0 : W-1$ **do**
    $\mathbf{C}_{(x)} \leftarrow z$ s.t. $\mathbf{S}^{\text{pred}}_{(z,x)} = k+1$ & $\mathbf{S}^{\text{pred}}_{(z-1,x)} \neq k+1$
    if $x = 0$ **then**
      $\mathbf{B}^{\text{pred}}_{(k,x)} \leftarrow \min \mathbf{C}_{(x)}$
      $b_{\text{prev}} \leftarrow \mathbf{B}^{\text{pred}}_{(k,x)}$
    else if $\mathbf{C}_{(x)}$ is not empty **then**
      $\mathbf{B}^{\text{pred}}_{(k,x)} \leftarrow \arg\min_{\mathbf{C}_{(x)}} |\mathbf{C}_{(x)} - b_{\text{prev}}|$
      $b_{\text{prev}} \leftarrow \mathbf{B}^{\text{pred}}_{(k,x)}$
    end if
  end for
  Interpolate $\mathbf{B}^{\text{pred}}_{(k)}$ for missing x values using b-spline
end for

5.4 Experimental result

In the experiment, we test and compare the segmentation performance of our proposed joint motion correction and 3D segmentation neural networks with several state-of-the-art conventional or deep learning methods.

5.4.1 Datasets

The methods are evaluated on three different datasets: DME dataset [50], AMD and Control dataset [36], and our own dataset collected by Jacobs Retina Center.

We use the DME dataset [50] as a benchmark of the proposed method to compare with the state of the art methods in literature. The DME dataset [50] is one of the most widely used public datasets in literature [140, 63, 139]. The dataset includes 10 macular centered OCT volumes for patients with DME imaged by Heidelberg Spectralis OCT system after motion correction. The resolution for each volume is $496 \times 768 \times 61$, with voxel size ranging from $3.87 \times 11.07 \times 118\mu$m to $3.87 \times 11.59 \times 128\mu$m. 11 selected B-scans out of 61 B-scans in each volume (in total 110 B-scans) are manually annotated with 8 segmentation boundaries ($K = 9$ classes) in the central region. We follow the training and test division in other papers [140, 63, 139], where the first 55 images from subject 1 to 5 are used for training, and the last 55 images from subject 6 to 10 are used for testing.

We then use the AMD and control dataset [36] to evaluate the influence of the motion correction network in 3D segmentation on OCT volumes with real motion artifacts. The AMD and control dataset [36] is a public dataset with 384 macular centered OCT volumes from 269 patients with AMD and 115 normal control subjects. The OCT volumes are imaged by the Bioptigen system and mostly has resolution $512 \times 1000 \times 100$, with some exceptions that has 82 B-scans. Manual annotations for 3 layer boundaries ($K = 4$ classes) are provided in a central circular region. Due to the different definition of the RPE-DC layer in AMD group [36], we only utilized normal control group for evaluation of the segmentation methods. The first 55 OCT volumes are used for training, and next 5 volumes are used for validation, and the last 55 volumes are used for testing.

Finally, the JRC dataset is used to compare the proposed method to OCT segmentation solutions clinically available to ophthalmologists in segmentation of retinal layers with various diseases. The JRC dataset contains 190 horizontal and vertical OCT volumes imaged with the Heidelberg Spectralis system [21], and 8 layers for 30 OCT volumes are manually corrected using the Heidelberg HEYEX software based on Heidelberg's segmentation result. The OCT volumes

without manual corrections are divided into 142 and 18 for training and validation using Heidelberg's segmentation as ground truth, and the 30 manually corrected volumes are divided into 15 and 15 for fine-tuning and testing. The resolution of the OCT volumes are $496 \times 512 \times 49$ with size $1.9 \times 5.8 \times 5.8$ mm^3. The dataset includes both normal subjects and patients with wet and dry AMD, nonproliferative diabetic retinopathy (NPDR), epi-retinal membrane (ERM), central retinal vein occlusion (CRVO), retinal detachment, macular hole, chorioretinopathy, and so on.

5.4.2 Simulated motion for DME dataset

For the AMD and control dataset and JRC dataset, we directly apply our motion correction approach on the original motion corrupted OCT volumes. Since the DME dataset [50] has been motion-corrected, we include simulated motion on the input OCT volumes to test the performance of our proposed motion correction approach. The simulated axial eye motion is generated using a similar method in [125], which is based on cumulative sum of Gaussian vector. We also verify the similarity of simulated motion and real eye motion by comparing their statistics in the AMD and control dataset and JRC dataset. Fig. 5.5 sub-figure (a) shows the histogram of motion amplitudes of the real and simulated motion vectors, and it can be observed that the real and simulated motion amplitudes follow a similar distribution. Fig. 5.5 sub-figure (b)-(e) respectively shows the normalized auto-correlation of 10 example motion vectors in the AMD and control dataset, JRC dataset, simulated motion, and Gaussian random vectors. The auto-correlation of real motion on both datasets are significantly different from Gaussian random vectors, and the auto-correlation of simulated motion resembles that of real motion.

5.4.3 Evaluation metrics

The classification error and the Dice loss are used to evaluate the pixel-wise performance of each segmentation algorithm. Specifically, we present the overall error, the error of retinal layers, the averaged Dice loss for all layers, and the Dice loss for each layer. Denoting the predicted binary segmentation map with $\mathbf{S}^{\text{pred}}$, ground truth segmentation with $\mathbf{S}^{\text{GT}}$, valid mask with $\mathbf{M}$, and element-wise product with $\odot$. The Dice loss for the kth layer can be obtained by one minus the

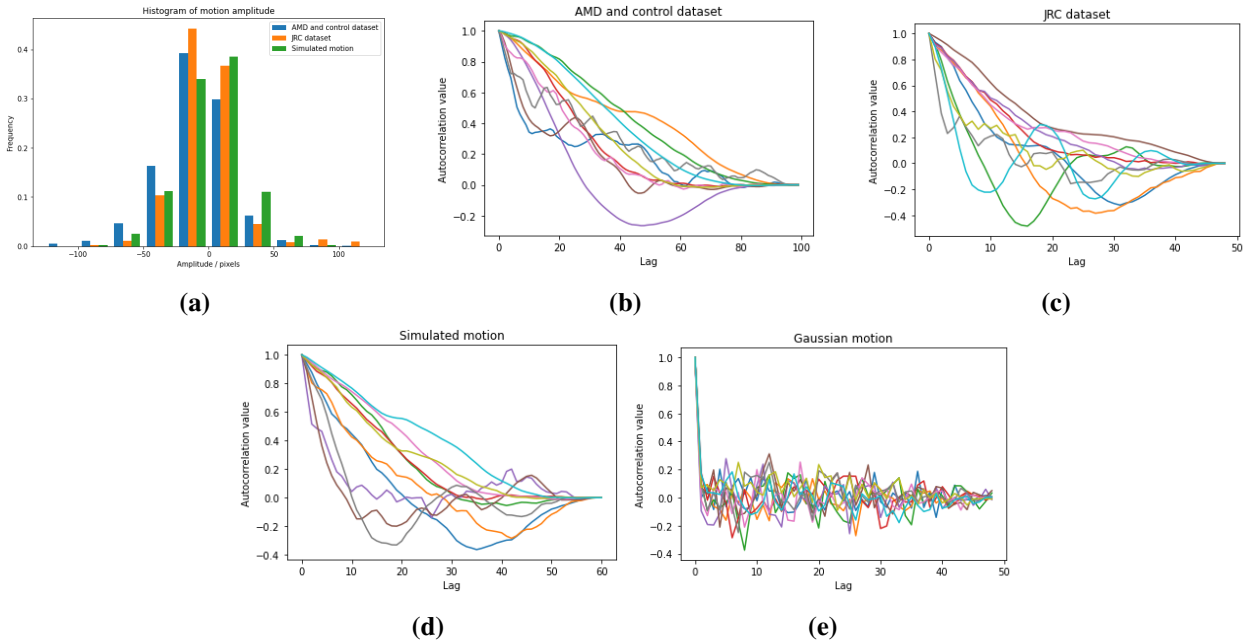


Figure 5.5. Statistics of real and simulated eye motion. (a) Histogram of motion amplitudes, (b)-(e) autocorrelation of 10 example motion vectors in the AMD and control dataset, JRC dataset, simulated motion, and Gausian random vectors.

Dice coefficient

$$\mathscr{L}_{\text{Dice},k}(\mathbf{S}^{\text{pred}},\mathbf{S}^{\text{GT}}) = 1 - \frac{2\sum \mathbf{S}_k^{\text{pred}} \odot \mathbf{S}_k^{\text{GT}} \odot \mathbf{M}}{\sum \mathbf{S}_k^{\text{pred}} \odot \mathbf{M} + \sum \mathbf{S}_k^{\text{GT}} \odot \mathbf{M}}, \tag{5.7}$$

and the averaged Dice loss for all retinal layers is

$$\mathscr{L}_{\text{Dice}}(\mathbf{S}^{\text{pred}},\mathbf{S}^{\text{GT}}) = \underset{k=1:K-2}{\text{mean}} \mathscr{L}_{\text{Dice},k}(\mathbf{S}^{\text{pred}},\mathbf{S}^{\text{GT}}). \tag{5.8}$$

The pixel-wise error is calculated on the valid region given by $\mathbf{M}$, and non-retinal regions ($k = 0$ or $K-1$) are merged into one class. The layer error is calculated based on retinal layers corresponding to class 1 to $K-2$ in the ground truth label.

$$\text{Error}(\mathbf{S}^{\text{pred}},\mathbf{S}^{\text{GT}}) = \frac{\sum(\mathbf{S}^{\text{pred}} \neq \mathbf{S}^{\text{GT}}) \odot \mathbf{M}}{\sum \mathbf{M}}. \tag{5.9}$$

After converting the pixel-wise predictions into layer boundaries using the proposed Algorithm 1, the mean absolute distance (MAD) is evaluated between the predicted and ground truth

boundaries in the annotated region masked by $\mathbf{M}^{\mathrm{b}} \in \mathbb{R}^{K \times W \times N}$,

$$\mathrm{MAD}(\mathbf{B}^{\mathrm{pred}}, \mathbf{B}^{\mathrm{GT}}) = \frac{\sum |\mathbf{B}^{\mathrm{pred}} - \mathbf{B}^{\mathrm{GT}}| \odot \mathbf{M}^{\mathrm{b}}}{\sum \mathbf{M}^{\mathrm{b}}}. \tag{5.10}$$

5.4.4 Implementation

In the experiment, we compare the performance of our proposed 3D segmentation with 7 B-scans input (center ±3 neighboring B-scans) and the 2D version using a single B-scan input with or without OCT motion correction network. We compare with several conventional methods by Chiu et al. [50] and Rathke et al. [137], the OCTExplorer software [134], as well as deep learning method U-Net [66], RelayNet [140], MGU-Net [139], and the network proposed by He et al. [63].

Our proposed motion correction and segmentation networks are implemented in PyTorch. The motion correction network utilizes the pre-trained model in [125]. On the DME dataset, the 2D network is first trained on the first 55 images with expert 1's annotation as ground truth, using batch size 4 for 200 epochs with an initial learning rate of 10^{-3} and decayed to 10^{-4} after 100 epochs, using Adam optimizer with weight decay of 10^{-4}. Since the dataset is sparsely annnotated, we use the prediction of the 2D network as pseudo-ground truth for B-scan slices without manual annotations to obtain dense label for the 3D OCT volume. The 3D segmentation network is then trained based on 3D labels with batch size 1. On the AMD and control dataset, the segmentation networks are trained using batch size 4 for 15 epochs with an initial learning rate of 10^{-3} and decayed to 10^{-4} after 10 epochs. We first pre-train the model on the JRC training set using Heidelberg segmentation as ground truth for 15 epochs with initial learning rate 10^{-3} and weight decay of 10^{-4}, and then fine-tune on 15 OCT volumes with manual labels for 20 epochs with learning rate 5×10^{-4} and 10 epochs with learning rate 10^{-4} using weight decay of 10^{-3}.

The method by Chiu et al. [50] uses the predicted layer boundaries provided in the DME dataset, and the method by Rathke et al. [137] uses the original implementation in Matlab. The OCTExplorer [134] software version 3.8.0 is used. We include both the pre-trained PyTorch model on the DME dataset provided by the original authors of RelayNet [140], and also include our

retrained model on all three datasets. The network by He et el. [63] uses the prediction results on the DME dataset provided by the authors. The U-Net [66] and MGU-Net [139] are trained in PyTorch using similar hyper parameters as our proposed segmentation network.

5.4.5 DME dataset

The qualitative results of segmentation on the original motion-corrected DME dataset are shown in Fig. 5.6, where the first group (a) shows the pixel-wise segmentation and group (b) shows the layer boundaries. Sub-figure (1) shows the input B-scan cropped in the labeled retinal region, sub-figure (2) shows the ground truth manual label, and sub-figures (3) to (10) show the segmentation result of different segmentation methods. The results demonstrate that the conventional methods by OCTExplorer [134] and Rathke et al. [137] could not accurately segment the retinal layers for the B-scan with DME. The method by Chiu et al. [50] produces more accurate segmentation, while the RNFL (in blue) is thinner than the ground truth, and the OPL (in yellow) does not follow the shape of lesions in the B-scan. For deep learning methods, the boundaries between each class of the RelayNet [140] prediction is noisier compared with the MGU-Net [139] and our proposed networks. MGU-Net [139] yields mis-classification of the OPL, and He et al. [63] yields discontinuities denoted by yellow arrows. The segmentation result of our proposed networks with 2D or 3D input are both visually continuous, and our 3D network result in lower Dice loss and MAD.

We also visualize the central slow B-scans in Fig. 5.7 to compare the 3D consistency of the MGU-Net and our proposed networks after applying simulated motion and our motion correction network. It can be observed that the simulated motion in sub-figure (a1) causes axial distortion to the slow B-scan, and it can be effectively corrected by our motion correction network in sub-figure (b1). Overall, our proposed 3D network in sub-figure (b4) after motion correction achieves the best consistency, compare to MGU-Net [139] in sub-figures (b2) and our 2D network in sub-figure (b3) at the location denoted by yellow arrow. It demonstrates the joint motion correction and 3D segmentation networks can improve the performance of 3D consistency.

Table 5.1. Comparison of pixel-wise label of different segmentation methods on DME test dataset [50] where the best and the second best are denoted by bold text and blue text, respectively.

					Dice Loss per Layer						
Data	Method	Error	Layer Error	Mean Dice Loss	RNFL	GCL-IPL	INL	OPL	ONL-ISM	ISE	OS-RPE
Original	Chiu et al. [50]	2.55%	13.46%	0.1616 (±0.063)	0.1490	0.1059	0.2439	0.2510	0.0694	0.1317	0.1805
	OCTExplorer [134]	9.64%	47.63%	0.5515 (±0.139)	0.5853	0.4200	0.6266	0.6625	0.2698	0.6453	0.6509
	Rathke et al. [137]	5.42%	29.82%	0.3279 (±0.090)	0.2935	0.2518	0.4101	0.4368	0.1687	0.4051	0.3290
	U-Net [66]	2.25%	10.63%	0.1360 (±0.062)	0.1131	0.0865	0.2441	0.2050	0.0585	0.1036	0.1414
	RelayNet [140]	3.18%	16.86%	0.1997 (±0.067)	0.2203	0.1412	0.2854	0.2689	0.0809	0.1732	0.2279
	RelayNet [140] (retrain)	2.46%	12.61%	0.1441 (±0.055)	0.1295	0.1081	0.2248	0.2298	0.0823	0.1043	0.1301
	MGU-Net [139]	2.13%	11.46%	0.1359 (±0.054)	0.1356	0.1038	0.2118	0.2152	0.0565	0.1032	0.1249
	He et al. [63]	2.18%	10.88%	0.1393 (±0.058)	**0.1116**	0.0792	0.1855	0.2240	0.0580	0.1245	0.1924
	Ours (2D)	1.95%	10.26%	0.1238 (±0.048)	0.1176	0.0864	0.1850	0.1946	0.0526	0.1004	0.1296
	Ours (3D)	**1.80%**	**9.75%**	**0.1155** (±0.045)	0.1137	**0.0732**	**0.1665**	**0.1847**	**0.0485**	**0.0998**	**0.1218**
Simulated Motion	Ours (2D)	1.92%	10.06%	0.1216 (±0.048)	0.1165	0.0860	0.1855	0.1931	0.0513	**0.0984**	**0.1205**
	Ours (3D)	2.21%	12.04%	0.1401 (±0.048)	0.1370	0.0917	0.1830	0.2004	0.0578	0.1348	0.1760
	Ours (2D, motion corrected)	1.92%	10.15%	0.1219 (±0.048)	**0.1158**	0.0855	0.1848	0.1947	0.0525	0.0990	0.1211
	Ours (3D, motion corrected)	**1.85%**	**10.03%**	**0.1189** (±0.045)	0.1165	**0.0770**	**0.1727**	**0.1855**	**0.0495**	0.1069	0.1239

Table 5.2. Comparison of segmentation boundaries of different segmentation methods on the DME test dataset [50] where the best and the second best are denoted by bold text and blue text, respectively.

			Mean Absolute Distance (MAD) per Layer							
Data	Method	MAD	ILM	RNFL	GCL-IPL	INL	OPL	ONL-ISM	ISE	OS-RPE
Original	Chiu et al. [50]	1.5723 (±0.475)	1.3146	1.6623	1.8948	2.1675	2.3025	0.9982	1.1213	1.1175
	OCTExplorer [134]	7.8175 (±0.705)	8.8859	8.4021	8.2109	7.8384	8.1318	6.7051	7.4995	6.8663
	Rathke et al. [137]	4.6272 (±1.381)	4.5134	5.7654	5.3873	5.9651	5.7223	2.3398	4.9436	2.3808
	U-Net [66]	1.7796 (±0.432)	2.0935	1.7927	2.2347	1.9086	2.3776	1.0859	1.3869	1.3569
	RelayNet [140]	2.0903 (±0.542)	1.6345	2.8419	2.4331	2.5968	2.5939	1.3682	1.6287	1.6250
	RelayNet [140] (retrain)	2.0310 (±0.459)	1.7375	2.3206	2.4817	2.3928	2.6642	1.3617	1.5098	1.7797
	MGU-Net [139]	1.4934 (±0.440)	1.5402	1.9179	1.8086	1.8004	2.0251	0.8981	1.0423	0.9149
	He et al. [63]	1.3190 (±0.351)	**1.0348**	1.4022	1.3791	1.7694	1.9076	0.7938	1.1446	1.1209
	Ours (2D)	1.7860 (±0.283)	1.7129	2.0226	1.8373	2.0534	2.2302	1.3568	1.5514	1.5232
	Ours (3D)	**1.1624** (±0.307)	1.4368	**1.2911**	**1.2020**	**1.3595**	**1.6210**	**0.7439**	**0.8648**	**0.7799**
Simulated Motion	Ours (2D)	1.8274 (±0.297)	1.9250	2.0946	1.8363	2.0508	2.2693	1.3620	1.5491	1.5324
	Ours (3D)	1.6090 (±0.263)	1.3448	1.7890	1.5729	1.8232	2.0311	1.1383	1.6132	1.5595
	Ours (2D, motion corrected)	1.9163 (±0.301)	1.8915	2.1713	1.9412	2.1898	2.3955	1.4781	1.6505	1.6123
	Ours (3D, motion corrected)	**1.2040** (±0.289)	**1.3398**	**1.3465**	**1.2598**	**1.4134**	**1.6813**	**0.8154**	**0.9264**	**0.8498**

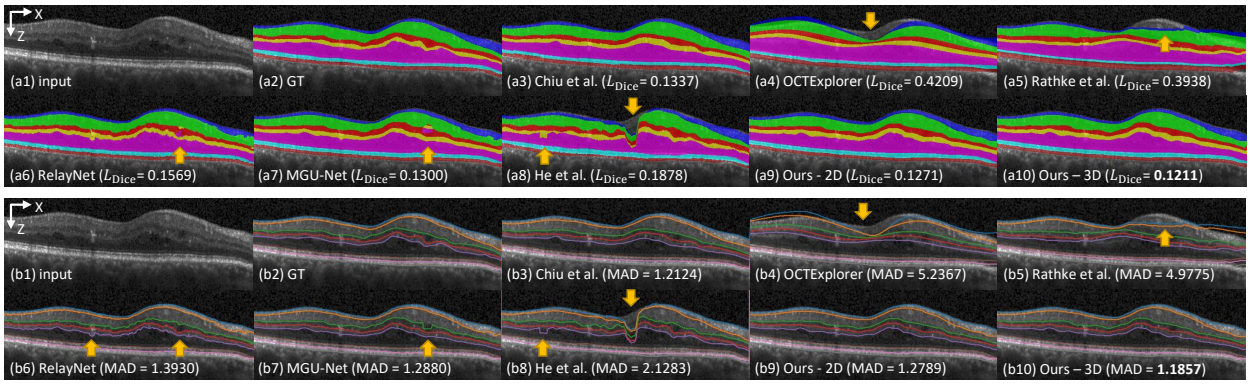


Figure 5.6. Qualitative results on the DME dataset [50]. Group (a) shows the pixel-wise prediction of each method with corresponding Dice loss, and group (b) shows the layer boundaries with mean absolute distance (MAD). (1) Input B-scan, (2) ground truth segmentation, (3) Chiu et al. [50], (4) Rathke et al. [137], (5) OCTExplorer [134], (6) RelayNet [140], (7) MGU-Net [139], (8) He et al. [63], (9) our proposed 2D network, (10) our proposed 3D network. Yellow arrows denote large segmentation errors.

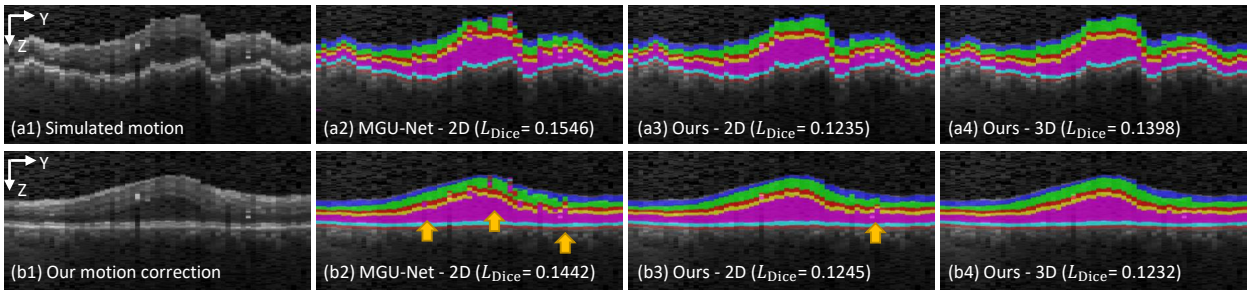


Figure 5.7. Qualitative comparison of 3D consistency on the DME dataset [50]. Group (a) shows results for OCT with simulate motion, group (b) shows results for motion-corrected OCT. (1) slow B-scan, (2)-(4) segmentation result of MGU-Net [139] and our proposed 2D or 3D network. Yellow arrows denote large segmentation errors.

Quantitative evaluation for pixel-wise accuracy on the DME dataset [50] is shown in Table 5.1. The raw network output without boundary detection post-processing described in Algorithm 1 is used in this evaluation for deep learning methods, and the pixel-wise labels are derived for conventional methods using the predicted boundaries. The best performance in each column for original OCT and OCT with simulated motion is denoted by bold text, and the second best is denoted by blue text. We only include our proposed methods in the experiment with simulated motion in order to evaluate the performance of our motion correction network. On the original input, our proposed 3D segmentation network achieves the lowest error of 1.80%, layer accuracy of 9.75%, and average Dice loss of 0.1155. The Dice loss of our 3D network is also the lowest in

the each retinal layer, except for the RNFL layer where it ranks as the third best result. For OCT with simulated motion, our proposed 3D segmentation network after motion correction achieves the lowest error and Dice coefficient, and note that the error of the 3D network increases without motion correction.

The mean average distance of layer boundaries are evaluated in Table 5.2. The proposed boundary detection algorithm is used to post-process the deep learning methods. Overall, the proposed 3D approach achieves the lowest average MAD at 1.1624 pixels, and the result by He et al. [63] achieves the second lowest average MAD at 1.3190 pixels. When comparing the the results on OCT with simulated motion, our 3D segmentation network with motion correction also achieves the lowest MAD with a improvement upon our 2D network and our 3D network without motion correction.

5.4.6 AMD and control dataset

We use the AMD and control dataset to evaluate the influence of the motion correction network on real motion corrupted OCT volumes. We visualize one example OCT volume in Fig. 5.8, where segmentation on the original 3D OCT volume is shown in group (1), and segmentation on the motion-corrected OCT volume is shown in group (2). The 3D OCT is shown in sub-figure (a), the segmentation surface of our 3D network is shown in sub-figure (b), and the partially annotated ground truth segmentation surface is shown in sub-figure (c). Sub-figures (d)-(f) show the central cross-section slow B-scan, our segmentation, and ground truth respectively, and sub-figures (g)-(i) show segmentation on the central fast B-scan. It can be observed that eye motion distorts the slow B-scan in sub-figures (a1), and causes jittered segmentation surface in (b1), (c1), (e1), and (f1). After applying the motion correction network, the segmentation surfaces are smooth along the slow-scanning axis and the slow B-scan provides better visualization.

The quantitative performance of several methods on the AMD and control dataset [36] are shown in Table 5.3, where our proposed 3D network after motion correction achieves the lowest error and Dice loss for each layer. It can also be observed that the performance of our 3D network

degrades without the motion correction network, demonstrating the importance of motion correction on dataset with real motion artifacts.

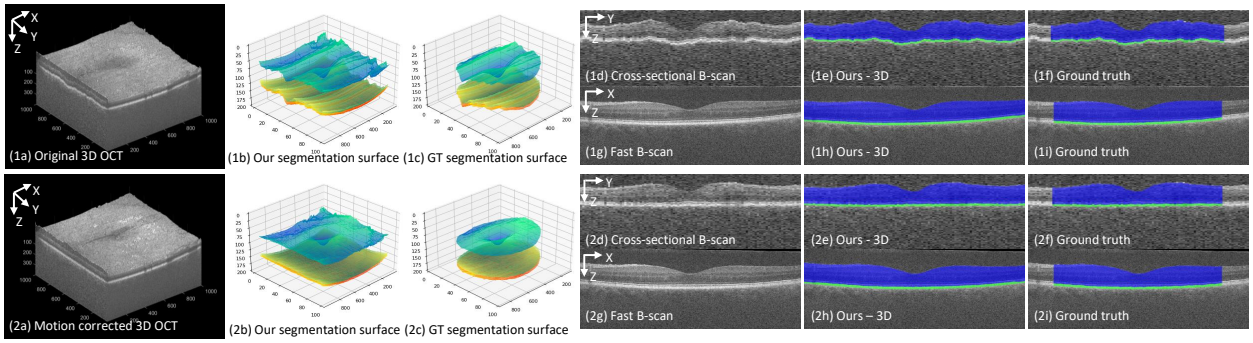


Figure 5.8. Qualitative results on the AMD and control dataset. Group 1 shows segmentation on the original OCT, group 2 shows segmentation on the motion-corrected OCT. (a) 3D OCT volume, (b) our segmentation surface, (c) ground truth (partially annotated) segmentation surface, (d)-(f) segmentation of the slow B-scan, (g)-(i) segmentation of the fast B-scan.

Table 5.3. Quantitative result of different methods on the AMD and control test dataset [36], where the best and the second best are denoted by bold and blue text, respectively.

			Dice Loss per Layer	
Method	Error	Layer Error	RNFL-OS	RPE
Rathke et al. [137]	1.21%	5.50%	0.0292	0.3078
U-Net [66]	0.83%	2.31%	0.0135	0.1039
RelayNet [140]	1.87%	2.86%	0.0172	0.1126
MGU-Net [139]	0.56%	2.03%	0.0111	0.1061
Ours (2D, original)	0.49%	1.82%	0.0093	0.1022
Ours (3D, original)	0.48%	1.86%	0.0092	0.0990
Ours (2D, motion corrected)	0.44%	1.61%	0.0083	0.0944
Ours (3D, motion corrected)	**0.41%**	**1.56%**	**0.0074**	**0.0892**

5.4.7 JRC dataset

We compare the proposed method to the clinically available solutions [134, 21] on the JRC dataset with various diseases. The qualitative results on the JRC dataset are shown in Fig. 5.9. Two examples with wet AMD and CRVO are illustrated in group (1) and group (2), respectively. Sub-figure (a) shows the 3D OCT volume with motion artifacts, sub-figure (b) shows the motion corrected 3D OCT volume, sub-figure (c) shows the slow B-scan with motion artifacts, sub-figure (d) shows the motion corrected B-scan, and sub-figure (e) shows a reference vertical B-scan imaged separately. Sub-figure (g) shows the segmentation of different methods on the fast B-scan, and

sub-figure (h) shows the segmentation of different methods on the motion corrected slow B-scan. The conventional methods by Rathke et al. [137], Heidelberg [21], and OCTExplorer [2] yield significant segmentation errors of the OPL and GLC-IPL in example (1), as denoted by pink and red arrows. These methods also produce large errors denoted by red circles in example (2) compared to the ground truth in sub-figure (g2-8) and sub-figure (h2-8). This is because the conventional methods rely on graph prior designed for normal eyes or diseases with mild deformations, which could not generalize well for various diseases with large deformations. Deep learning methods including the U-Net [66], RelayNet [140], MGU-Net [139], and our proposed 3D network could produce segmentation with higher similarity to ground truth for both examples compared with conventional methods. However, the segmentation results of U-Net [66], RelayNet [140], and MGU-Net [139] yield mis-classifications as denoted by red circles in sub-figures (g2-4) to (g2-6), and they also yield and 3D inconsistency as denoted by red circles in sub-figures (h2-4) to (h2-6). The proposed 3D network produces more accurate segmentation in the fast B-scans, and also yields better 3D consistency in the slow B-scans for both examples.

We also visualize the 3D segmentation surfaces of Heidelberg [21], MGU-Net [139], our 3D network, and ground truth in Fig. 5.10. One quarter of the OCT is cut to show the cross-section of the segmentation surfaces. It could be observed that Heidelberg and MGU-Net produce segmentation errors denoted in red circles, and the proposed method is the most similar to the ground truth.

The quantitative results are presented in Table 5.4, where the proposed 3D segmentation network is compared with different segmentation methods. Since the manual annotation is performed based on Heidelberg's segmentation, the quantitative results would be biased towards Heidelberg. Therefore we also report the error evaluated only on manually corrected areas, where at least 4 out of 8 layers in the ground truth differ from Heidelberg's segmentation. The OCT volumes in the test set are divided into three categories, including normal, moderate deformation, and severe deformation. The percentage of area that is manually corrected is 1.02% for normal, 11.41% for moderate, and 21.47% for severe deformation. On the entire test set, our proposed 3D segmentation

Table 5.4. Quantitative result of different segmentation methods on the JRC test dataset, where the best and the second best are denoted by bold text and blue text, respectively. Note that ground truth labels are manually corrected based on Heidelberg's result.

					Dice Loss per Layer						
Data	Method	Error	Layer Error	Dice Loss	RNFL	GCL-IPL	INL	OPL	ONL-ISM	ISE	OS-RPE
All	Rathke et al. [137]	4.51%	27.09%	0.2820 (±0.076)	0.2755	0.2190	0.3228	0.2976	0.2240	0.4376	0.1976
	OCTExplorer [134]	3.78%	22.11%	0.2291 (±0.063)	0.1983	0.1342	0.2195	0.3148	0.1657	0.2858	0.2855
	Heidelberg [21]	1.74%	10.96%	0.1070 (±0.028)	0.0972	0.1152	0.1519	0.1378	0.0807	0.0693	0.0966
	U-Net [66]	1.58%	9.34%	0.0930 (±0.024)	0.0855	0.0760	0.1217	0.1291	0.0566	0.0825	0.0998
	RelayNet [140]	1.32%	7.93%	0.0816 (±0.024)	0.0739	0.0675	0.1134	0.1220	0.0523	0.0686	0.0736
	MGU-Net [139]	1.25%	7.54%	0.0788 (±0.024)	**0.0702**	**0.0635**	0.1069	0.1215	**0.0508**	0.0715	0.0668
	Ours (2D, original)	1.24%	7.45%	0.0771 (±0.023)	0.0694	0.0647	**0.1043**	**0.1183**	0.0516	0.0657	0.0656
	Ours (3D, original)	1.49%	8.22%	0.0902 (±0.026)	0.0850	0.0838	0.1236	0.1329	0.0568	0.0659	0.0836
	Ours (3D, motion corrected)	**1.22%**	**7.13%**	**0.0766** (±0.024)	0.0709	0.0652	0.1055	0.1211	**0.0508**	**0.0628**	**0.0596**
Normal	Rathke et al. [137]	3.22%	19.67%	0.2077 (±0.086)	0.2267	0.1399	0.2007	0.2027	0.1819	0.3955	0.1063
	OCTExplorer [134]	2.99%	16.50%	0.1835 (±0.056)	0.2003	0.0928	0.1542	0.2499	0.1446	0.2637	0.1788
	Heidelberg [21]	**0.22%**	**1.26%**	**0.0138** (±0.004)	**0.0165**	**0.0100**	**0.0167**	**0.0213**	**0.0096**	**0.0097**	**0.0126**
	U-Net [66]	1.02%	5.38%	0.0612 (±0.021)	0.0505	0.0409	0.1010	0.0820	0.0360	0.0542	0.0639
	RelayNet [140]	0.87%	4.64%	0.0531 (±0.023)	0.0472	0.0404	0.0978	0.0753	0.0265	0.0359	0.0484
	MGU-Net [139]	0.81%	4.67%	0.0511 (±0.019)	0.0478	0.0408	0.0815	0.0767	0.0269	0.0372	0.0471
	Ours (2D, original)	0.78%	4.53%	0.0493 (±0.019)	0.0467	0.0404	0.0799	0.0742	0.0262	0.0341	0.0433
	Ours (3D, original)	0.90%	4.93%	0.0558 (±0.020)	0.0513	0.0449	0.0870	0.0837	0.0310	0.0370	0.0558
	Ours (3D, motion corrected)	1.22%	7.13%	0.0766 (±0.024)	0.0709	0.0652	0.1055	0.1211	0.0508	0.0628	0.0596
Moderate deformation	Rathke et al. [137]	3.48%	21.60%	0.2292 (±0.082)	0.2307	0.1734	0.2604	0.2380	0.1853	0.3983	0.1182
	OCTExplorer [134]	3.24%	19.01%	0.2025 (±0.058)	0.1919	0.1128	0.1883	0.2844	0.1429	0.2635	0.2338
	Heidelberg [21]	**0.55%**	**3.28%**	**0.0340** (±0.008)	**0.0431**	**0.0315**	**0.0410**	**0.0427**	**0.0200**	**0.0279**	**0.0315**
	U-Net [66]	1.14%	7.20%	0.0722 (±0.021)	0.0708	0.0597	0.0995	0.1011	0.0386	0.0638	0.0719
	RelayNet [140]	1.06%	6.51%	0.0675 (±0.022)	0.0677	0.0586	0.0998	0.0963	0.0350	0.0519	0.0631
	MGU-Net [139]	1.04%	6.43%	0.0669 (±0.021)	0.0653	0.0552	0.0985	0.0960	0.0348	0.0562	0.0625
	Ours (2D, original)	1.00%	6.12%	0.0642 (±0.021)	0.0647	0.0569	0.0959	0.0936	0.0347	0.0490	0.0547
	Ours (3D, original)	1.25%	6.87%	0.0771 (±0.025)	0.0788	0.0734	0.1131	0.1079	0.0401	0.0485	0.0776
	Ours (3D, motion corrected)	1.01%	6.09%	0.0654 (±0.023)	0.0662	0.0583	0.0997	0.0983	0.0355	0.0464	0.0531
Severe deformation	Rathke et al. [137]	7.48%	42.36%	0.4326 (±0.080)	0.3800	0.3618	0.5138	0.4904	0.3448	0.5596	0.3778
	OCTExplorer [134]	5.38%	31.43%	0.3097 (±0.088)	0.2088	0.2044	0.3142	0.4210	0.2404	0.3549	0.4245
	Heidelberg [21]	5.18%	32.01%	0.3116 (±0.090)	0.2378	0.3694	0.4445	0.4143	0.2772	0.1977	0.2406
	U-Net [66]	2.84%	15.82%	0.1545 (±0.033)	0.1246	0.1315	0.1754	0.2246	0.1237	0.1492	0.1526
	RelayNet [140]	2.14%	12.55%	0.1289 (±0.037)	0.0945	0.1022	0.1477	0.2088	0.1200	0.1313	0.0979
	MGU-Net [139]	1.95%	11.30%	0.1213 (±0.040)	0.0870	0.0946	0.1354	0.2068	0.1139	0.1316	0.0796
	Ours (2D, original)	2.00%	11.71%	0.1213 (±0.036)	**0.0859**	0.0956	0.1324	0.1997	0.1186	0.1276	0.0895
	Ours (3D, original)	2.34%	12.69%	0.1388 (±0.037)	0.1083	0.1289	0.1610	0.2197	0.1230	0.1294	0.1012
	Ours (3D, motion corrected)	**1.89%**	**10.53%**	**0.1157** (±0.037)	0.0872	**0.0933**	**0.1264**	**0.1965**	**0.1105**	**0.1229**	**0.0734**
Manually corrected area	Rathke et al. [137]	6.85%	38.90%	0.4002 (±0.075)	0.3697	0.3332	0.4707	0.4529	0.3159	0.5242	0.3350
	OCTExplorer [134]	5.10%	29.87%	0.2970 (±0.079)	0.2162	0.2025	0.3020	0.4072	0.2235	0.3368	0.3905
	Heidelberg [21]	4.72%	29.22%	0.2855 (±0.077)	0.2458	0.3379	0.4000	0.3727	0.2386	0.1829	0.2207
	U-Net [66]	2.56%	14.89%	0.1445 (±0.033)	0.1212	0.1268	0.1724	0.2114	0.1115	0.1356	0.1328
	RelayNet [140]	2.10%	12.41%	0.1268 (±0.036)	0.0990	0.1056	0.1528	0.2032	0.1097	0.1210	0.0961
	MGU-Net [139]	1.93%	11.37%	0.1203 (±0.038)	0.0927	0.0976	0.1419	0.2013	0.1044	0.1235	0.0808
	Ours (2D, original)	1.96%	11.60%	0.1195 (±0.035)	**0.0915**	0.0993	0.1388	0.1956	0.1077	0.1163	0.0874
	Ours (3D, original)	2.27%	12.45%	0.1351 (±0.036)	0.1125	0.1279	0.1632	0.2116	0.1120	0.1178	0.1005
	Ours (3D, motion corrected)	**1.85%**	**10.52%**	**0.1137** (±0.036)	0.0922	**0.0964**	**0.1324**	**0.1911**	**0.1006**	**0.1107**	**0.0728**
Normal	Rathke et al. [137]	3.72%	22.40%	0.2211 (±0.083)	0.2283	0.1519	0.2352	0.2259	0.1863	0.3990	0.1208
	OCTExplorer [134]	3.30%	17.98%	0.1968 (±0.052)	0.1690	0.1436	0.1806	0.2765	0.1511	0.2759	0.1808
	Heidelberg [21]	1.28%	7.14%	0.0794 (±0.030)	0.0592	0.0721	**0.1136**	0.1362	0.0564	0.0501	0.0685
	U-Net [66]	1.60%	7.56%	0.0914 (±0.036)	0.0561	0.0749	0.1614	0.1286	0.0666	0.0672	0.0850
	RelayNet [140]	1.28%	6.80%	0.0785 (±0.035)	0.0475	0.0687	0.1397	0.1239	0.0488	0.0526	0.0686
	MGU-Net [139]	1.18%	6.65%	0.0745 (±0.031)	0.0462	0.0662	0.1184	0.1253	0.0461	0.0534	0.0662
	Ours (2D, original)	1.17%	6.65%	0.0740 (±0.031)	0.0463	0.0677	0.1170	0.1273	0.0481	0.0501	0.0618
	Ours (3D, original)	1.21%	6.60%	0.0755 (±0.032)	0.0477	0.0682	0.1184	0.1299	0.0504	0.0479	0.0662
	Ours (3D, motion corrected)	**1.10%**	**6.11%**	**0.0696** (±0.032)	**0.0446**	**0.0654**	0.1145	**0.1218**	**0.0446**	**0.0407**	**0.0552**
Moderate deformation	Rathke et al. [137]	4.32%	25.80%	0.2771 (±0.079)	0.2770	0.2254	0.3183	0.3080	0.2224	0.4259	0.1624
	OCTExplorer [134]	3.78%	22.65%	0.2379 (±0.058)	0.2111	0.1660	0.2460	0.3390	0.1626	0.2723	0.2681
	Heidelberg [21]	1.96%	12.12%	0.1263 (±0.036)	0.1590	0.1407	0.1649	0.1612	0.0767	0.0893	0.0923
	U-Net [66]	1.66%	10.77%	0.1065 (±0.029)	0.1029	0.0993	0.1442	0.1521	0.0659	0.0882	0.0927
	RelayNet [140]	1.62%	10.14%	0.1042 (±0.032)	0.0963	0.1007	0.1491	0.1549	0.0658	0.0778	0.0844
	MGU-Net [139]	1.58%	9.95%	0.1029 (±0.031)	0.0958	**0.0936**	0.1454	0.1524	0.0645	0.0853	0.0830
	Ours (2D, original)	1.51%	9.51%	0.0987 (±0.032)	0.0947	0.0964	**0.1404**	0.1506	0.0637	0.0713	0.0736
	Ours (3D, original)	1.75%	10.07%	0.1103 (±0.034)	0.1083	0.1107	0.1561	0.1610	0.0691	0.0725	0.0944
	Ours (3D, motion corrected)	**1.48%**	**9.06%**	**0.0966** (±0.033)	**0.0946**	0.0949	0.1411	**0.1481**	**0.0622**	**0.0672**	**0.0683**
Severe deformation	Rathke et al. [137]	8.34%	45.96%	0.4695 (±0.081)	0.4221	0.3908	0.5499	0.5421	0.3766	0.5901	0.4150
	OCTExplorer [134]	5.88%	33.89%	0.3316 (±0.092)	0.2210	0.2219	0.3317	0.4501	0.2655	0.3803	0.4504
	Heidelberg [21]	6.34%	38.47%	0.3746 (±0.104)	0.2981	0.4481	0.5204	0.4947	0.3414	0.2420	0.2777
	U-Net [66]	3.08%	17.21%	0.1675 (±0.037)	0.1325	0.1415	0.1851	0.2487	0.1461	0.1702	0.1482
	RelayNet [140]	2.39%	13.77%	0.1425 (±0.043)	0.1030	0.1094	0.1550	0.2330	0.1436	0.1522	0.1011
	MGU-Net [139]	2.16%	12.27%	0.1337 (±0.047)	0.0940	0.1009	0.1413	0.2314	0.1359	0.1516	0.0805
	Ours (2D, original)	2.23%	12.84%	0.1342 (±0.042)	**0.0926**	0.1020	0.1390	0.2228	0.1421	0.1483	0.0930
	Ours (3D, original)	2.60%	13.87%	0.1526 (±0.042)	0.1178	0.1385	0.1680	0.2432	0.1457	0.1512	0.1035
	Ours (3D, motion corrected)	**2.07%**	**11.43%**	**0.1268** (±0.043)	0.0939	**0.0986**	**0.1293**	**0.2175**	**0.1309**	**0.1427**	**0.0748**

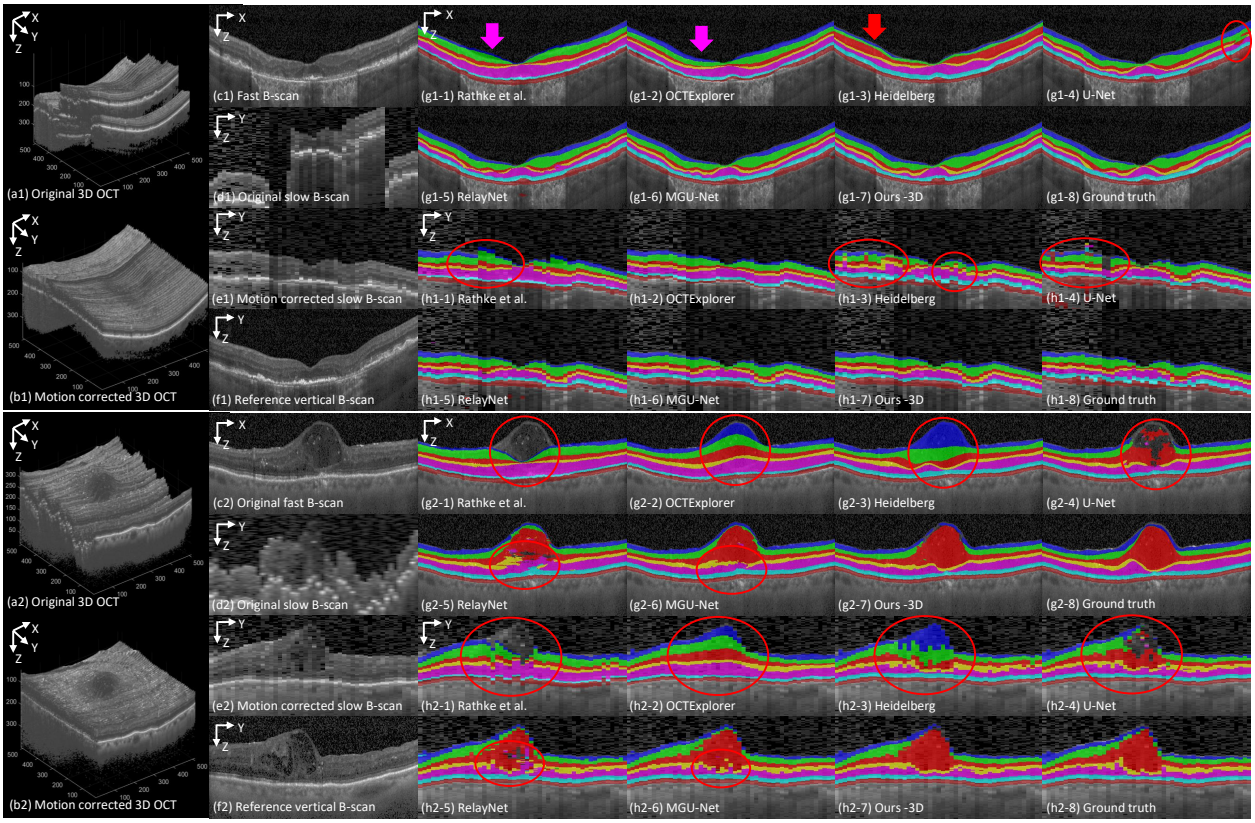


Figure 5.9. Qualitative results on the JRC dataset. Group (1) and (2) show two examples. (a) Original 3D OCT volume, (b) motion corrected 3D OCT volume, (c) fast B-scan, (d) original slow B-scan, (e) motion-corrected slow B-scan, (f) reference vertical B-scan, (g) segmentation on the fast B-scan using different methods, (h) segmentation on the slow B-scan using different methods. Pink arrows denote large error in OPL, red arrow denotes large error in GCL-IPL, and red circle denotes large segmentation error.

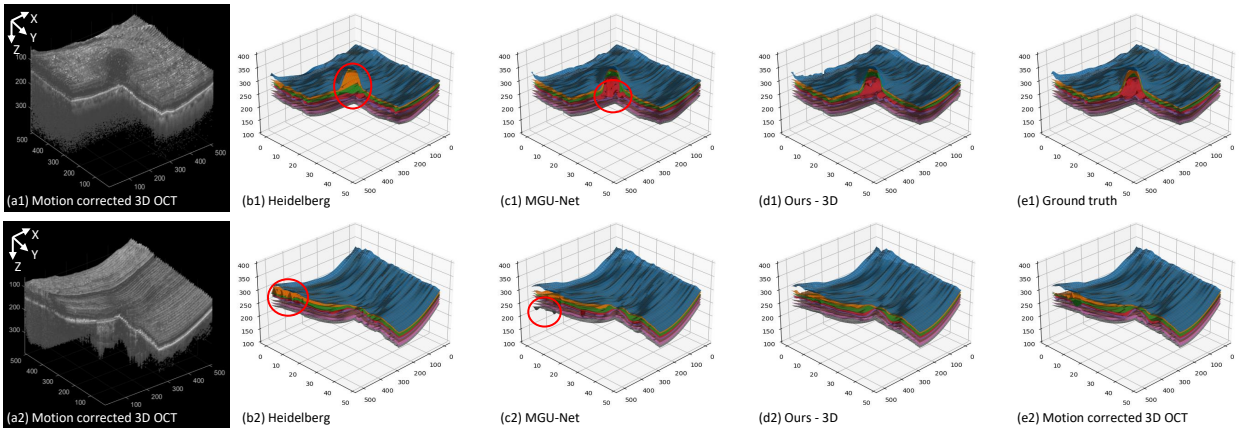


Figure 5.10. Visualization of segmentation result in 3D on the JRC dataset. Group (1) and (2) show two examples. (a) Motion corrected 3D OCT volume, (b)-(e) segmentation surfaces of Heidelberg, MGU-Net, our 3D network, and manual annotated ground truth. Red circle denotes large segmentation error.

network with motion correction yields the lowest error and Dice loss on average. When divided into three categories based on diseases, Heidelberg achieves the lowest error for normal and moderate deformation due to the bias of the ground truth. However, our proposed method outperforms Heidelberg segmentation by a large margin for the category of severe deformation, decreasing the layer error from 32.01% to 10.53%, and the average Dice loss from 0.3116 to 0.1157. For evaluation on the manually corrected area, the proposed 3D segmentation network with motion correction achieves the best performance overall and in each category of diseases. The results demonstrate a significant advantage of the combined motion correction and 3D segmentation network for a clinical dataset with various diseases.

5.5 Conclusion

In this chapter, we proposed to combine motion correction and 3D OCT layer segmentation, which led to promising improvement upon existing 2D segmentation methods with or without motion correction. Experimental results demonstrated that the motion correction is essential to apply 3D segmentation, and combining motion correction with 3D segmentation achieved the best performance for three datasets compared to conventional and deep learning state-of-the-art methods. Specifically, the proposed network demonstrated a significant advantage over clinically available segmentation solutions for severe diseases.

In future work, the segmentation network can be extended to support segmentation of retinal fluid and other lesions. The proposed segmentation method could be used to generate more accurate OCT-A projection images, promote the analysis of layer thickness and vessel density, which is beneficial for diagnosing and monitoring retinal and systemic diseases.

Chapter 5, in part, has been submitted for publication of the material as it may appear in IEEE International Conference on Image Processing 2022 (ICIP), Y. Wang, C. Galang, W. R. Freeman, T. Q. Nguyen, C. An, IEEE, 2021. The dissertation author is the primary author of this paper.

Chapter 5, in part is currently being prepared for submission for publication of the material.

Y. Wang, C. Galang, A. Warter, A. Heinke, D. G. Bartsch, W. R. Freeman, T. Q. Nguyen, C. An. The dissertation author is the primary author of this paper.

Chapter 6

Conclusion and Future Work

Retinal diseases including age-related macular degeneration, diabetes retinopathy, and vascular occlusions are important causes of vision loss and have systemic implications for millions of patients. The role of imaging and image processing is crucial in retinal diseases. As treatments advance, it is important to be able to scientifically analyze and interpret a large amount of information procured from different areas on the retina, measured by different instruments, and evaluate retinal structure and function over time and in response to therapies. It is increasingly difficult for any individual specialist or reading center to reliably review the multiple types of imaging available in a patient, nor align and overlay these imaging modalities to analyze retinal structure and function and determine correlations and predictive value of these tests to clinical outcomes. The ability to accurately align and overlay multimodal data, correct artifacts, and enhance data, segment and extract critical information from data with novel image processing techniques will promote the development of understanding and treatment of retinal disease.

This dissertation proposes multiple deep learning-based algorithms for retinal image processing, including multimodal retinal image registration, OCT motion correction, and OCT retinal layer segmentation:

- In Chapter 2, we proposed a content-adaptive weakly supervised deep learning framework for multimodal retinal image registration. The proposed method consists of three neural networks for vessel segmentation, feature detection and description, and outlier rejection. The proposed method demonstrated significant improvement compared with conventional

methods in both registration accuracy and robustness to diseases and poor imaging quality. A preliminary version of the method proposed in this chapter was published in the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020) [72], and an enhanced version was published in the IEEE Transactions on Image Processing [129].

- In Chapter 3, we presented a comprehensive overview of the existing evaluation metrics for multimodal retinal image registration, and compared the Pearson correlation coefficient between the ophthalmologists' subjective grade and several commonly used objective evaluation metrics. We found that the soft Dice coefficient with the segmentation method in [31] achieved the highest correlation with the subjective grades. The study in this chapter was published in the IEEE Access [142].

- In Chapter 4, we proposed a deep learning approach to correct axial and coronal motion artifacts in OCT. Experimental results demonstrated that the proposed method was able to correct large motion while recovering the retinal curvature, achieving significant improvements compared to the conventional methods. The preliminary axial correction network was published in the 2021 IEEE International Conference on Image Processing (ICIP 2021) [125], and an enhanced version was submitted to the IEEE Transactions on Image Processing.

- In Chapter 5, we proposed a joint OCT motion correction and a 3D retinal layer segmentation network that utilizes 3D contextual information. The proposed networks achieved the best performance for three datasets compared to conventional and deep learning state-of-the-art methods. The preliminary study was submitted to the 2022 IEEE International Conference on Image Processing (ICIP 2022), and an enhanced version was to be submitted to the IEEE Transactions on Image Processing.

In future work, the proposed research could be further extended in the following aspects:

- The proposed multimodal retinal image registration framework could be optimized for image pairs with larger field of view difference, in order to register standard FOV images with ultra-wide angle images.

- The proposed multimodal retinal image registration framework could be integrated into a visualization software to help ophthalmologists observe co-localized multimodal images. The aligned images can be visualized by checkerboard overlay or alpha-blending.

- The proposed multimodal retinal image registration framework could be applied to register functional tests, such as microperimetry, with structural tests, such as OCT. Accurate registration would enable the development of structural-functional prediction algorithms based on co-localized images.

- An alternative OCT motion correction network could be developed to flatten the BM surface of the retina, when it's only required to generate OCT-A projection images. As the retinal curvature in the axial direction would not influence OCT-A projection, the motion correction problem could be made easier by flattening the retina.

- The proposed joint motion correction and OCT segmentation network could be applied to improve OCT-A projections especially for various diseases. It would lead to more accurate visualization of vessels in each retinal layer and facilitate the analysis of OCT-A vascular changes under various diseases.

- When applied for OCT-A segmentation, the proposed OCT segmentation network could be improved by adding 3D OCT-A to the input. Combining OCT and OCT-A information would help the segmentation of BM layer, which would be difficult using OCT alone in diseases like CNV.

- The proposed OCT segmentation network could be extended to segment layer-specific features, retinal fluid, and other lesions, which could potentially generate topology guaranteed segmentation and enable automatic disease identification.

The proposed retinal image processing via the novel deep learning approaches will not only promote diagnosis and analysis of retinal diseases, but will also provide insights on observing systematic diseases through the retina, including diabetes, cardiovascular disease, and preclinical

Alzheimer Disease. The proposed deep learning approaches for multimodal retinal image processing would build a connection between ophthalmology and image processing literature, and the findings may provide a good insight for researchers who investigate retinal image registration, retinal image segmentation, and retinal disease detection.

Bibliography

[1] A. London, I. Benhar, and M. Schwartz, "The retina as a window to the brain—from eye research to cns disorders," *Nature Reviews Neurology*, vol. 9, no. 1, p. 44, 2013.

[2] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE reviews in biomedical engineering*, vol. 3, pp. 169–208, 2010.

[3] T. J. MacGillivray, E. Trucco, J. R. Cameron, B. Dhillon, J. G. Houston, and E. J. R. Van Beek, "Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions," *The British journal of radiology*, vol. 87, no. 1040, p. 20130832, 2014.

[4] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, and G. Van Den Driessche, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[5] J. Yim, R. Chopra, T. Spitz, J. Winkens, A. Obika, C. Kelly, H. Askham, M. Lukic, J. Huemer, K. Fasler, and G. Moraes, "Predicting conversion to wet age-related macular degeneration using deep learning," *Nature Medicine*, vol. 26, no. 6, pp. 892–899, 2020.

[6] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Deep retinal image understanding," in *International conference on medical image computing and computer-assisted intervention*, pp. 140–148, Springer, 2016.

[7] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "Dunet: A deformable network for retinal vessel segmentation," *Knowledge-Based Systems*, vol. 178, pp. 149–162, 2019.

[8] L. Li, M. Verma, Y. Nakashima, H. Nagahara, and R. Kawasaki, "Iternet: Retinal image segmentation utilizing structural redundancy in vessel networks," 2019.

[9] J. A. Lee, P. Liu, J. Cheng, and H. Fu, "A deep step pattern representation for multimodal retinal image registration," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5077–5086, 2019.

[10] H. Kolb, "How the retina works: Much of the construction of an image takes place in the retina itself through the use of specialized neural circuits," *American scientist*, vol. 91, no. 1, pp. 28–35, 2003.

[11] X. Zhang, J. B. Saaddine, C.-F. Chou, M. F. Cotch, Y. J. Cheng, L. S. Geiss, E. W. Gregg, A. L. Albright, B. E. Klein, and R. Klein, "Prevalence of diabetic retinopathy in the united states, 2005-2008," *Jama*, vol. 304, no. 6, pp. 649–656, 2010.

[12] D. S. Friedman, B. J. O'Colmain, B. Munoz, S. C. Tomany, C. McCarty, P. De Jong, B. Nemesure, P. Mitchell, J. Kempen, and N. Congdon, "Prevalence of age-related macular degeneration in the united states," *Arch ophthalmol*, vol. 122, no. 4, pp. 564–572, 2004.

[13] M. M. Abdull, C. Chandler, and C. Gilbert, "Glaucoma,"the silent thief of sight": patients' perspectives and health seeking behaviour in bauchi, northern nigeria," *BMC ophthalmology*, vol. 16, no. 1, pp. 1–9, 2016.

[14] T. Y. Wong, A. Shankar, R. Klein, B. E. Klein, and L. D. Hubbard, "Prospective cohort study of retinal vessel diameters and risk of hypertension," *bmj*, vol. 329, no. 7457, p. 79, 2004.

[15] A. Uchida, J. A. Pillai, R. Bermel, A. Bonner-Jackson, A. Rae-Grant, H. Fernandez, J. Bena, S. E. Jones, J. B. Leverenz, S. K. Srivastava, and J. P. Ehlers, "Outer retinal assessment using spectral-domain optical coherence tomography in patients with alzheimer's and parkinson's disease," *Investigative ophthalmology & visual science*, vol. 59, no. 7, pp. 2768–2777, 2018.

[16] A. Arrigo, M. Teussink, E. Aragona, F. Bandello, and M. B. Parodi, "Multicolor imaging to detect different subtypes of retinal microaneurysms in diabetic retinopathy," *Eye*, vol. 35, no. 1, pp. 277–281, 2021.

[17] I. Kozak, D.-U. Bartsch, L. Cheng, and W. R. Freeman, "In vivo histology of cotton-wool spots using high-resolution optical coherence tomography," *American journal of ophthalmology*, vol. 141, no. 4, pp. 748–750, 2006.

[18] A. Ly, L. Nivison-Smith, N. Assaad, and M. Kalloniatis, "Infrared reflectance imaging in age-related macular degeneration," *Ophthalmic and Physiological Optics*, vol. 36, no. 3, pp. 303–316, 2016.

[19] S. Schmitz-Valckenberg, F. G. Holz, A. C. Bird, and R. F. Spaide, "Fundus autofluorescence imaging: review and perspectives," *Retina*, vol. 28, no. 3, pp. 385–409, 2008.

[20] M. G. Bittencourt, M. Hassan, M. S. Halim, R. Afridi, N. V. Nguyen, C. Plaza, A. N. Tran, M. I. Ahmed, Q. D. Nguyen, and Y. J. Sepah, "Blue light versus green light fundus autofluorescence in normal subjects and in patients with retinochoroidopathy secondary to retinal and uveitic diseases," *Journal of ophthalmic inflammation and infection*, vol. 9, no. 1, pp. 1–9, 2019.

[21] M. M. Teussink, S. Donner, T. Otto, K. Williams, and A. Tafreshi, "State-of-the-art commercial spectral-domain and swept-source oct technologies and their clinical applications in ophthalmology citation," 2019.

[22] D. A. Salz and A. J. Witkin, "Imaging in diabetic retinopathy," *Middle East African journal of ophthalmology*, vol. 22, no. 2, p. 145, 2015.

[23] T. E. De Carlo, A. Romano, N. K. Waheed, and J. S. Duker, "A review of optical coherence tomography angiography (octa)," *International journal of retina and vitreous*, vol. 1, no. 1, p. 5, 2015.

[24] H. R. Novotny and D. L. Alvis, "A method of photographing fluorescence in circulating blood in the human retina," *Circulation*, vol. 24, no. 1, pp. 82–86, 1961.

[25] L. A. Yannuzzi, J. S. Slakter, J. A. Sorenson, D. R. Guyer, and D. A. Orlock, "Digital indocyanine green videoangiography and choroidal neovascularization.," *Retina (Philadelphia, Pa.)*, vol. 12, no. 3, pp. 191–223, 1992.

[26] J. Pitcher III and J. Hau, "Cscr. diagnosis and treatment," *Review of Ophthalmology*, 2014.

[27] G. Yang, C. V. Stewart, M. Sofka, and C.-L. Tsai, "Registration of challenging image pairs: Initialization, estimation, and decision," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 11, pp. 1973–1989, 2007.

[28] J. Chen, J. Tian, N. Lee, J. Zheng, R. T. Smith, and A. F. Laine, "A partial intensity invariant feature descriptor for multimodal retinal image registration," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1707–1718, 2010.

[29] G. Wang, Z. Wang, Y. Chen, and W. Zhao, "Robust point matching method for multimodal retinal image registration," *Biomedical Signal Processing and Control*, vol. 19, pp. 68–76, 2015.

[30] Z. Li, F. Huang, J. Zhang, B. Dashtbozorg, S. Abbasi-Sureshjani, Y. Sun, X. Long, Q. Yu, B. ter Haar Romeny, and T. Tan, "Multi-modal and multi-vendor retina image registration," *Biomedical optics express*, vol. 9, no. 2, pp. 410–422, 2018.

[31] J. Zhang, C. An, J. Dai, M. Amador, D.-U. Bartsch, S. Borooah, W. R. Freeman, and T. Q. Nguyen, "Joint vessel segmentation and deformable registration on multi-modal retinal images based on style transfer," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 839–843, IEEE, 2019.

[32] P. C. Cattin, H. Bay, L. Van Gool, and G. Székely, "Retina mosaicing using local features," in *International conference on medical image computing and computer-assisted intervention*, pp. 185–192, Springer, 2006.

[33] S. Lee, M. D. Abràmoff, and J. M. Reinhardt, "Retinal image mosaicing using the radial distortion correction model," in *Medical Imaging 2008: Image Processing*, vol. 6914, p. 691435, International Society for Optics and Photonics, 2008.

[34] S. Lee, M. D. Abramoff, and J. M. Reinhardt, "Retinal atlas statistics from color fundus images," in *Medical Imaging 2010: Image Processing*, vol. 7623, p. 762310, International Society for Optics and Photonics, 2010.

[35] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *International conference on medical image computing and computer-assisted intervention*, pp. 130–137, Springer, 1998.

[36] S. Farsiu, S. J. Chiu, R. V. O'Connell, F. A. Folgar, E. Yuan, J. A. Izatt, C. A. Toth, and A.-R. E. D. S. . A. S. D. O. C. T. S. Group, "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography," *Ophthalmology*, vol. 121, no. 1, pp. 162–172, 2014.

[37] M. Niemeijer, M. K. Garvin, B. van Ginneken, M. Sonka, and M. D. Abramoff, "Vessel segmentation in 3d spectral oct scans of the retina," in *Medical Imaging 2008: Image Processing*, vol. 6914, p. 69141R, International Society for Optics and Photonics, 2008.

[38] J. Xu, D. Tolliver, H. Ishikawa, G. Wollstein, and J. S. Schuman, "3d oct retinal vessel segmentation based on boosting learning," in *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany*, pp. 179–182, Springer, 2009.

[39] Z. Hu, M. Niemeijer, M. D. Abramoff, K. Lee, and M. K. Garvin, "Automated segmentation of 3-d spectral oct retinal blood vessels by neural canal opening false positive suppression," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 33–40, Springer, 2010.

[40] J. Wu, B. S. Gerendas, S. M. Waldstein, G. Langs, C. Simader, and U. Schmidt-Erfurth, "Stable registration of pathological 3d-oct scans using retinal vessels," 2014.

[41] R. G. Sayegh, C. G. Kiss, C. Simader, J. Kroisamer, A. Montuoro, T. J. Mittermüller, M. Azhary, M. Bolz, D. P. Kreil, and U. Schmidt-Erfurth, "A systematic correlation of morphology and function using spectral domain optical coherence tomography and microperimetry in patients with geographic atrophy," *British Journal of Ophthalmology*, vol. 98, no. 8, pp. 1050–1055, 2014.

[42] C. A. Toth, V. Tai, M. Pistilli, S. J. Chiu, K. P. Winter, E. Daniel, J. E. Grunwald, G. J. Jaffe, D. F. Martin, G.-s. Ying, S. Farsiu, and M. G. Maguire, "Distribution of oct features within areas of macular atrophy or scar after 2 years of anti-vegf treatment for neovascular amd in catt," *Ophthalmology Retina*, vol. 3, no. 4, pp. 316–325, 2019.

[43] M. Nagpal, J. Khandelwal, R. Juneja, and N. Mehrotra, "Correlation of optical coherence tomography angiography and microperimetry (mp3) features in wet age-related macular degeneration," *Indian journal of ophthalmology*, vol. 66, no. 12, p. 1790, 2018.

[44] N. Ritter, R. Owens, J. Cooper, R. H. Eikelboom, and P. P. Van Saarloos, "Registration of stereo and temporal images of the retina," *IEEE Transactions on medical imaging*, vol. 18, no. 5, pp. 404–418, 1999.

[45] T. Chanwimaluang, G. Fan, and S. R. Fransen, "Hybrid retinal image registration," *IEEE transactions on information technology in biomedicine*, vol. 10, no. 1, pp. 129–142, 2006.

[46] L. Sánchez Brea, D. Andrade De Jesus, M. F. Shirazi, M. Pircher, T. van Walsum, and S. Klein, "Review on retrospective procedures to correct retinal motion artefacts in oct imaging," *Applied Sciences*, vol. 9, no. 13, p. 2700, 2019.

[47] B. Potsaid, I. Gorczynska, V. J. Srinivasan, Y. Chen, J. Jiang, A. Cable, and J. G. Fujimoto, "Ultrahigh speed spectral/fourier domain oct ophthalmic imaging at 70,000 to 312,500 axial scans per second," *Optics express*, vol. 16, no. 19, pp. 15149–15169, 2008.

[48] A. Montuoro, J. Wu, S. Waldstein, B. Gerendas, G. Langs, C. Simader, and U. Schmidt-Erfurth, "Motion artefact correction in retinal optical coherence tomography using local symmetry," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 130–137, Springer, 2014.

[49] J. Lauermann, A. Woetzel, M. Treder, M. Alnawaiseh, C. Clemens, N. Eter, and F. Alten, "Prevalences of segmentation errors and motion artifacts in oct-angiography differ among retinal diseases," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 256, no. 10, pp. 1807–1816, 2018.

[50] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," *Biomedical optics express*, vol. 6, no. 4, pp. 1172–1194, 2015.

[51] R. D. Ferguson, D. X. Hammer, L. A. Paunescu, S. Beaton, and J. S. Schuman, "Tracking optical coherence tomography," *Optics letters*, vol. 29, no. 18, pp. 2139–2141, 2004.

[52] Y. K. Tao, S. Farsiu, and J. A. Izatt, "Interlaced spectrally encoded confocal scanning laser ophthalmoscopy and spectral domain optical coherence tomography," *Biomedical optics express*, vol. 1, no. 2, pp. 431–440, 2010.

[53] M. F. Kraus, B. Potsaid, M. A. Mayer, R. Bock, B. Baumann, J. J. Liu, J. Hornegger, and J. G. Fujimoto, "Motion correction in optical coherence tomography volumes on a per a-scan basis using orthogonal scan patterns," *Biomedical optics express*, vol. 3, no. 6, pp. 1182–1199, 2012.

[54] B. Antony, M. D. Abramoff, L. Tang, W. D. Ramdas, J. R. Vingerling, N. M. Jansonius, K. Lee, Y. H. Kwon, M. Sonka, and M. K. Garvin, "Automated 3-d method for the correction of axial artifacts in spectral-domain optical coherence tomography images," *Biomedical optics express*, vol. 2, no. 8, pp. 2403–2416, 2011.

[55] J. Xu, H. Ishikawa, G. Wollstein, L. Kagemann, and J. S. Schuman, "Alignment of 3-d optical coherence tomography scans to correct eye movement using a particle filtering," *IEEE transactions on medical imaging*, vol. 31, no. 7, pp. 1337–1345, 2012.

[56] D. C. Fernández, H. M. Salinas, and C. A. Puliafito, "Automated detection of retinal layer structures on optical coherence tomography images," *Optics express*, vol. 13, no. 25, pp. 10200–10216, 2005.

[57] Q. Yang, C. A. Reisman, Z. Wang, Y. Fukuma, M. Hangai, N. Yoshimura, A. Tomidokoro, M. Araie, A. S. Raza, D. C. Hood, and K. Chan, "Automated layer segmentation of macular oct images using dual-scale gradient information," *Optics express*, vol. 18, no. 20, pp. 21293–21307, 2010.

[58] A. Carass, A. Lang, M. Hauser, P. A. Calabresi, H. S. Ying, and J. L. Prince, "Multiple-object geometric deformable model for segmentation of macular oct," *Biomedical optics express*, vol. 5, no. 4, pp. 1062–1074, 2014.

[59] J. Novosel, G. Thepass, H. G. Lemij, J. F. de Boer, K. A. Vermeer, and L. J. van Vliet, "Loosely coupled level sets for simultaneous 3d retinal layer segmentation in optical coherence tomography," *Medical image analysis*, vol. 26, no. 1, pp. 146–158, 2015.

[60] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in oct images of non-exudative amd patients using deep learning and graph search," *Biomedical optics express*, vol. 8, no. 5, pp. 2732–2744, 2017.

[61] C. S. Lee, A. J. Tyring, N. P. Deruyter, Y. Wu, A. Rokem, and A. Y. Lee, "Deep-learning based, automated segmentation of macular edema in optical coherence tomography," *Biomedical optics express*, vol. 8, no. 7, pp. 3440–3448, 2017.

[62] Y. He, A. Carass, Y. Yun, C. Zhao, B. M. Jedynak, S. D. Solomon, S. Saidha, P. A. Calabresi, and J. L. Prince, "Towards topological correct segmentation of macular oct from cascaded fcns," in *Fetal, Infant and Ophthalmic Medical Image Analysis*, pp. 202–209, Springer, 2017.

[63] Y. He, A. Carass, Y. Liu, B. M. Jedynak, S. D. Solomon, S. Saidha, P. A. Calabresi, and J. L. Prince, "Fully convolutional boundary regression for retina oct segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 120–128, Springer, 2019.

[64] Y. He, A. Carass, Y. Liu, B. M. Jedynak, S. D. Solomon, S. Saidha, P. A. Calabresi, and J. L. Prince, "Deep learning based topology guaranteed surface and mme segmentation of multiple sclerosis subjects from retinal oct," *Biomedical optics express*, vol. 10, no. 10, pp. 5042–5058, 2019.

[65] M. Pekala, N. Joshi, T. A. Liu, N. M. Bressler, D. C. DeBuc, and P. Burlina, "Deep learning based retinal oct segmentation," *Computers in biology and medicine*, vol. 114, p. 103445, 2019.

[66] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[67] M. K. Garvin, M. D. Abràmoff, R. Kardon, S. R. Russell, X. Wu, and M. Sonka, "Intraretinal layer segmentation of macular optical coherence tomography images using optimal 3-d graph search," *IEEE transactions on medical imaging*, vol. 27, no. 10, pp. 1495–1505, 2008.

[68] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6148–6157, 2017.

[69] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical image analysis*, vol. 52, pp. 128–143, 2019.

[70] H. Zhang, X. Liu, G. Wang, Y. Chen, and W. Zhao, "An automated point set registration framework for multimodal retinal image," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2857–2862, IEEE, 2018.

[71] J. A. Lee, J. Cheng, B. H. Lee, E. P. Ong, G. Xu, D. W. K. Wong, J. Liu, A. Laude, and T. H. Lim, "A low-dimensional step pattern analysis algorithm with application to multimodal retinal image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1046–1053, 2015.

[72] Y. Wang, J. Zhang, C. An, M. Amador, D.-U. Bartsch, W. R. Freeman, and T. Q. Nguyen, "A segmentation based robust deep learning framework for multimodal retinal image registration," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1369–1373, 2020.

[73] J. Zhang, B. Dashtbozorg, E. Bekkers, J. P. Pluim, R. Duits, and B. M. ter Haar Romeny, "Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores," *IEEE transactions on medical imaging*, vol. 35, no. 12, pp. 2631–2644, 2016.

[74] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey vision conference*, vol. 15, pp. 10–5244, Citeseer, 1988.

[75] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005.

[76] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, pp. 1150–1157, Ieee, 1999.

[77] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*, pp. 404–417, Springer, 2006.

[78] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *European Conference on Computer Vision*, pp. 467–483, Springer, 2016.

[79] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," in *Advances in Neural Information Processing Systems*, pp. 2414–2422, 2016.

[80] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018.

[81] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[82] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American statistical association*, vol. 79, no. 388, pp. 871–880, 1984.

[83] O. Chum and J. Matas, "Matching with prosac-progressive sample consensus," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 220–226, IEEE, 2005.

[84] O. Chum and J. Matas, "Optimal randomized ransac," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1472–1482, 2008.

[85] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, "Usac: a universal framework for random sample consensus," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 2022–2038, 2012.

[86] C.-L. Tsai, C.-Y. Li, G. Yang, and K.-S. Lin, "The edge-driven dual-bootstrap iterative closest point algorithm for registration of multimodal fluorescein angiogram sequence," *IEEE transactions on medical imaging*, vol. 29, no. 3, pp. 636–649, 2009.

[87] Z. Ghassabi, J. Shanbehzadeh, A. Sedaghat, and E. Fatemizadeh, "An efficient approach for robust multimodal retinal image registration based on ur-sift features and piifd descriptors," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 25, 2013.

[88] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac-differentiable ransac for camera localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6684–6692, 2017.

[89] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to find good correspondences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2666–2674, 2018.

[90] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2758–2766, 2015.

[91] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462–2470, 2017.

[92] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018.

[93] J. Hur and S. Roth, "Iterative residual refinement for joint optical flow and occlusion estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5754–5763, 2019.

[94] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: a learning framework for deformable medical image registration," *IEEE transactions on medical imaging*, 2019.

[95] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 204–212, Springer, 2017.

[96] D. Mahapatra, S. Sedai, and R. Garnavi, "Elastic registration of medical images with gans," *arXiv preprint arXiv:1805.02369*, 2018.

[97] D. Mahapatra, B. Antony, S. Sedai, and R. Garnavi, "Deformable medical image registration using generative adversarial networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 1449–1453, IEEE, 2018.

[98] H. Su, V. Jampani, D. Sun, O. Gallo, E. Learned-Miller, and J. Kautz, "Pixel-adaptive convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11166–11175, 2019.

[99] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE transactions on signal processing*, vol. 49, no. 12, pp. 3136–3144, 2001.

[100] A. Cifor, L. Risser, D. Chung, E. M. Anderson, and J. A. Schnabel, "Hybrid feature-based diffeomorphic registration for tumor tracking in 2-d liver ultrasound images," *IEEE transactions on medical imaging*, vol. 32, no. 9, pp. 1647–1656, 2013.

[101] A. Wong, D. A. Clausi, and P. Fieguth, "Cpol: Complex phase order likelihood as a similarity measure for mr–ct registration," *Medical image analysis*, vol. 14, no. 1, pp. 50–57, 2010.

[102] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, pp. 694–711, Springer, 2016.

[103] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE transactions on medical imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[104] N.Mayer, E.Ilg, P.Häusser, P.Fischer, D.Cremers, A.Dosovitskiy, and T.Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.

[105] S. H. M. Alipour, H. Rabbani, and M. R. Akhlaghi, "Diabetic retinopathy grading by digital curvelet transform," *Computational and Mathematical Methods in Medicine*, vol. 2012.

[106] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[107] D. Freedman, R. Pisani, and R. Purves, *Statistics: Fourth International Student Edition*. W.W. Norton & Company, 2007.

[108] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[109] R. A. Fisher, "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika*, vol. 10, no. 4, pp. 507–521, 1915.

[110] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, M. Brady, and J. A. Schnabel, "Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Medical image analysis*, vol. 16, no. 7, pp. 1423–1435, 2012.

[111] IBM Corp., "Ibm spss statistics for windows."

[112] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, and J. G. Fujimoto, "Optical coherence tomography," *science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[113] A. Baghaie, Z. Yu, and R. M. D'Souza, "Involuntary eye motion correction in retinal optical coherence tomography: Hardware or software solution?," *Medical image analysis*, vol. 37, pp. 129–145, 2017.

[114] M. Bengs, N. Gessert, M. Schlüter, and A. Schlaefer, "Spatio-temporal deep learning methods for motion estimation using 4d oct image data," *International journal of computer assisted radiology and surgery*, vol. 15, no. 6, pp. 943–952, 2020.

[115] A. Li, C. Du, and Y. Pan, "Deep-learning-based motion correction in optical coherence tomography angiography," *Journal of Biophotonics*, vol. 14, no. 12, p. e202100097, 2021.

[116] Y. Chen, Y.-J. Hong, S. Makita, and Y. Yasuno, "Three-dimensional eye motion correction by lissajous scan optical coherence tomography," *Biomedical optics express*, vol. 8, no. 3, pp. 1783–1802, 2017.

[117] S. Y. Ksenofontov, P. A. Shilyagin, D. A. Terpelov, V. M. Gelikonov, and G. V. Gelikonov, "Numerical method for axial motion artifact correction in retinal spectral-domain optical coherence tomography," *Frontiers of Optoelectronics*, pp. 1–9, 2020.

[118] E. Gibson, M. Young, M. V. Sarunic, and M. F. Beg, "Optic nerve head registration via hemispherical surface and volume registration," *IEEE transactions on biomedical engineering*, vol. 57, no. 10, pp. 2592–2595, 2010.

[119] M. Niemeijer, K. Lee, M. K. Garvin, M. D. Abràmoff, and M. Sonka, "Registration of 3d spectral oct volumes combining icp with a graph-based approach," in *Medical Imaging 2012: Image Processing*, vol. 8314, p. 83141A, International Society for Optics and Photonics, 2012.

[120] M. F. Kraus, J. J. Liu, J. Schottenhamml, C.-L. Chen, A. Budai, L. Branchini, T. Ko, H. Ishikawa, G. Wollstein, J. Schuman, and J. Duker, "Quantitative 3d-oct motion correction with tilt and illumination correction, robust similarity measure and regularization," *Biomedical optics express*, vol. 5, no. 8, pp. 2591–2613, 2014.

[121] D. Gaucher, A. Erginay, A. Lecleire-Collet, B. Haouchine, M. Puech, S.-Y. Cohen, P. Massin, and A. Gaudric, "Dome-shaped macula in eyes with myopic posterior staphyloma," *American journal of ophthalmology*, vol. 145, no. 5, pp. 909–914, 2008.

[122] U. C. Park, D. J. Ma, W. H. Ghim, and H. G. Yu, "Influence of the foveal curvature on myopic macular complications," *Scientific reports*, vol. 9, no. 1, pp. 1–8, 2019.

[123] H. Fu, Y. Xu, D. W. K. Wong, and J. Liu, "Eye movement correction for 3d oct volume by using saliency and center bias constraint," in *2016 IEEE Region 10 Conference (TENCON)*, pp. 1536–1539, IEEE, 2016.

[124] A. Abdolmanafi, L. Duong, N. Dahdah, and F. Cheriet, "Intra-slice motion correction of intravascular oct images using deep features," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 931–941, 2018.

[125] Y. Wang, A. Warter, M. Cavichini-Cordeiro, W. R. Freeman, D.-U. G. Bartsch, T. Q. Nguyen, and C. An, "Learning to correct axial motion in oct for 3d retinal imaging," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 126–130, 2021.

[126] R. Rocholz, M. Teussink, R. Dolz-Marco, C. Holzhey, J. Dechent, A. Tafreshi, and S. Schulz, "Spectralis optical coherence tomography angiography (octa): principles and clinical applications," 2018.

[127] D. Huang, Y. Jia, S. S. Gao, B. Lumbroso, and M. Rispoli, "Optical coherence tomography angiography using the optovue device," *Oct Angiography in Retinal and Macular Diseases*, vol. 56, pp. 6–12, 2016.

[128] J. Yu and R. Ramamoorthi, "Robust video stabilization by optimization in cnn weight space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3800–3808, 2019.

[129] Y. Wang, J. Zhang, M. Cavichini, D.-U. G. Bartsch, W. R. Freeman, T. Q. Nguyen, and C. An, "Robust content-adaptive global registration for multimodal retinal images using weakly supervised deep-learning framework," vol. 30, pp. 3167–3178, IEEE, 2021.

[130] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3766–3778, 2013.

[131] J. G. Fujimoto, W. Drexler, J. S. Schuman, and C. K. Hitzenberger, "Optical coherence tomography (oct) in ophthalmology: introduction.," *Optics express*, vol. 17, no. 5, pp. 3978–3979, 2009.

[132] S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, and S. Farsiu, "Automatic segmentation of seven retinal layers in sdoct images congruent with expert manual segmentation," *Optics express*, vol. 18, no. 18, pp. 19413–19428, 2010.

[133] Y. He, A. Carass, Y. Liu, B. M. Jedynak, S. D. Solomon, S. Saidha, P. A. Calabresi, and J. L. Prince, "Structured layer surface segmentation for retina oct using fully convolutional regression networks," *Medical image analysis*, vol. 68, p. 101856, 2021.

[134] M. K. Garvin, M. D. Abràmoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka, "Automated 3-d intraretinal layer segmentation of macular spectral-domain optical coherence tomography images," *IEEE transactions on medical imaging*, vol. 28, no. 9, pp. 1436–1447, 2009.

[135] H. Bogunović, F. Venhuizen, S. Klimscha, S. Apostolopoulos, A. Bab-Hadiashar, U. Bagci, M. F. Beg, L. Bekalo, Q. Chen, C. Ciller, and K. Gopinath, "Retouch: the retinal oct fluid detection and segmentation benchmark and challenge," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1858–1874, 2019.

[136] D. Lu, M. Heisler, S. Lee, G. Ding, M. V. Sarunic, and M. F. Beg, "Retinal fluid segmentation and detection in optical coherence tomography images using fully convolutional neural network," *arXiv preprint arXiv:1710.04778*, 2017.

[137] F. Rathke, S. Schmidt, and C. Schnörr, "Probabilistic intra-retinal layer segmentation in 3-d oct images using global shape regularization," *Medical image analysis*, vol. 18, no. 5, pp. 781–794, 2014.

[138] S. J. Chiu, C. A. Toth, C. B. Rickman, J. A. Izatt, and S. Farsiu, "Automatic segmentation of closed-contour features in ophthalmic images using graph theory and dynamic programming," *Biomedical optics express*, vol. 3, no. 5, pp. 1127–1140, 2012.

[139] J. Li, P. Jin, J. Zhu, H. Zou, X. Xu, M. Tang, M. Zhou, Y. Gan, J. He, Y. Ling, and Y. Su, "Multi-scale gcn-assisted two-stage network for joint segmentation of retinal layers and discs in peripapillary oct images," *Biomedical Optics Express*, vol. 12, no. 4, pp. 2204–2220, 2021.

[140] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical optics express*, vol. 8, no. 8, pp. 3627–3642, 2017.

[141] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 433–442, 2019.

[142] Y. Wang, J. Zhang, M. Cavichini, D.-U. G. Bartsch, W. R. Freeman, T. Q. Nguyen, and C. An, "Study on correlation between subjective and objective metrics for multimodal retinal image registration," *IEEE Access*, vol. 8, pp. 190897–190905, 2020.