UNIVERSITY OF CALIFORNIA SAN DIEGO

Multi-Modal Retinal Image Registration via Deep Neural Networks

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

in

Electrical Engineering (Signal and Image Processing)

by

Junkang Zhang

Committee in charge:

Professor Truong Q. Nguyen, Chair
Professor Cheolhong An
Professor Pamela C. Cosman
Professor William R. Freeman
Professor David J. Kriegman

2022

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

University of California San Diego

2022

DEDICATION

This thesis is dedicated to my parents for their endless love and unconditional support.

This thesis is also dedicated to the lovely lady who suddenly showed up in my life during the pandemic and keeps me company through the days.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, my deepest gratitude should be given to my advisor Professor Truong Q. Nguyen for his all time support throughout my PhD study and research. His patience and knowledge have guided me to approach the end of the program which would be unreachable without him. In addition, I am also deeply grateful to Professor Cheolhong An who has been providing me with invaluable guidance and feedbacks using his expertise throughout all my PhD research process.

Besides, I am very grateful to Professor William R. Freeman for his crucial expert knowledge which helped my research task. I am also very thankful to Professor Pamela C. Cosman and Professor David J. Kriegman for supervising me through all my PhD exams.

Many thanks should also be given to the doctors and research fellows in Jacobs Retina Center who have been collaborating with me and collecting data for my research project. Finally, my thanks should go to the colleague students and visiting scholars in Video Processing Lab for the time we spent together.

Chapter 2, in full, is a reprint of the material as it appears in IEEE Transactions on Image Processing, 2021. Junkang Zhang; Yiqian Wang; Ji Dai; Melina Cavichini-Cordeiro; Dirk-Uwe G. Bartsch; William R. Freeman; Truong Q. Nguyen; Cheolhong An, IEEE, 2021. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, has been submitted for publication of the material as it may appear in IEEE Transactions on Image Processing, 2022, Junkang Zhang; Yiqian Wang; Fritz Gerald P. Kalaw; Melina Cavichini-Cordeiro; Dirk-Uwe G. Bartsch; William R. Freeman; Truong Q. Nguyen; Cheolhong An, IEEE, 2022. The dissertation author was the primary investigator and author of this material.

Chapter 4, in part, is currently being prepared for submission for publication of the material. Junkang Zhang; Yiqian Wang; Fritz Gerald P. Kalaw; Melina Cavichini-Cordeiro; Dirk-Uwe G. Bartsch; William R. Freeman; Truong Q. Nguyen; Cheolhong An. The dissertation author was the primary investigator and author of this material.

VITA

2014	B. E. in Automation, Hohai University, Changzhou, China
2017	M. E. in Pattern Recognition and Intelligent Systems, Southeast University, Nanjing, China
2022	Ph. D. in Electrical Engineering (Signal and Image Processing), University of California San Diego

PUBLICATIONS

Junkang Zhang, Yiqian Wang, Fritz Gerald P. Kalaw, Melina Cavichini-Cordeiro, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Cheolhong An. 3D Eyeball Shape Estimation for Ultra-Widefield and Narrow-Angle Retinal Image Alignment. To be submitted to *ICASSP 2023*.

Junkang Zhang, Yiqian Wang, Fritz Gerald P. Kalaw, Melina Cavichini-Cordeiro, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Cheolhong An. Multimodal Global Registration between Ultra-Widefield and Narrow Angle Retinal Images via Distortion Correction Network. Submitted to *IEEE Transactions on Image Processing*.

Cheolhong An, Yiqian Wang, Junkang Zhang, and Truong Q. Nguyen. Self-Supervised Rigid Registration for Multimodal Retinal Images. *IEEE Transactions on Image Processing*, vol. 31, pp. 5733-5747, 2022.

Junkang Zhang, Yiqian Wang, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Cheolhong An. Perspective Distortion Correction for Multi-Modal Registration between Ultra-Widefield and Narrow-Angle Retinal Images. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2021, pp. 4086-4091.

Junkang Zhang, Yiqian Wang, Ji Dai, Melina Cavichini, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Cheolhong An. Two-Step Registration on Multi-Modal Retinal Images via Deep Neural Networks. *IEEE Transactions on Image Processing*, vol. 31, pp. 823-838, 2021.

Yiqian Wang, Junkang Zhang, Melina Cavichini, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Cheolhong An. Robust Content-Adaptive Global Registration for Multi-modal Retinal Images Using Weakly Supervised Deep-Learning Framework. *IEEE Transactions on Image Processing*, vol. 30, pp. 3167-3178, 2021.

Yiqian Wang, Junkang Zhang, Melina Cavichini, Dirk-Uwe G. Bartsch, William R. Freeman, Truong Q. Nguyen, and Cheolhong An. Study on Correlation Between Subjective and Objective Metrics for Multimodal Retinal Image Registration. *IEEE Access*, vol. 8, pp. 190897-190905, 2020.

Ji Dai, Shiwei Jin, Junkang Zhang, and Truong Q. Nguyen. Boosting Feature Matching Accuracy With Pairwise Affine Estimation. *IEEE Transactions on Image Processing*, vol. 29, pp. 8278-8291, 2020.

Yiqian Wang, Junkang Zhang, Cheolhong An, Melina Cavichini, Mahima Jhingan, Manuel J. Amador-Patarroyo, Christopher P. Long, Dirk-Uwe G. Bartsch, William R. Freeman, and Truong Q. Nguyen. A Segmentation based Robust Deep Learning Framework for Multimodal Retinal Image Registration. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 1369-1373.

Junkang Zhang, Cheolhong An, Ji Dai, Manuel Amador, Dirk-Uwe Bartsch, Shyamanga Borooah, William R. Freeman, and Truong Q. Nguyen. Joint Vessel Segmentation and Deformable Registration on Multi-Modal Retinal Images based on Style Transfer. *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 839-843.

Ji Dai, Junkang Zhang, and Truong Q. Nguyen. Explicit Learning of Feature Orientation Estimation. *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4245-4249.

Junkang Zhang, Cheolhong An, Truong Nguyen. Deep Joint Demosaicing and Super Resolution on High Resolution Bayer Sensor Data. *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2018, pp. 619–623.

ABSTRACT OF THE DISSERTATION

Multi-Modal Retinal Image Registration via Deep Neural Networks

by

Junkang Zhang

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

University of California San Diego, 2022

Professor Truong Q. Nguyen, Chair

Multi-modal retinal images provide complementary anatomical information at various resolutions, color wavelengths, and fields of view. Aligning multi-modal images will establish a comprehensive view of the retina and benefit the screening and diagnosis of eye diseases. However, the inconsistent anatomical patterns across modalities create outliers in feature matching, and the lack of retinal boundaries may also fool the intensity-based alignment metrics, both of which will influence the alignment qualities. Besides, the varying distortion levels across Ultra-Widefield (UWF) and Narrow-Angle (NA) images, due to different camera parameters, will cause large alignment errors in global transformation.

In addressing the issue of inconsistent patterns, we use retinal vasculature as a common

signal for alignment. First, we build a two-step coarse-to-fine registration pipeline fully based on deep neural networks. The coarse alignment step estimates a global transformation via vessel segmentation, feature detection and description, and outlier rejection. While the fine alignment step corrects the remaining misalignment through deformable registration. In addition, we propose an unsupervised learning scheme based on style transfer to jointly train the networks for vessel segmentation and deformable registration. Finally, we also introduce Monogenical Phase signal as an alternative guidance in training the deformable registration network.

Then, to deal with the issue of various distortion levels across UWF and NA modalities, we propose a distortion correction function to create images with similar distortion levels. Based on the assumptions of spherical eyeball shape and fixed UWF camera pose, the function reprojects the UWF pixels by an estimated correction camera with similar parameters as the NA camera. Besides, we incorporate the function into the coarse alignment networks which will simultaneously optimize the correction camera pose and refine the global alignment results.

Moreover, to further reduce misalignment from the UWF-to-NA global registration, we estimate a 3D dense scene for the UWF pixels to represent a more flexible eyeball shape. Both the scene and the NA camera parameters are iteratively optimized to reduce the alignment error between the 3D-to-2D reprojected images and the original ones, which is also concatenated with the coarse alignment networks with distortion correction function.

1 Introduction

Retinal imaging plays an important role in ophthalmological diagnosis and treatment. Since modern imaging techniques can image retina in multiple modalities that demonstrate various levels of anatomical information, aligning multi-modal images can further provide a comprehensive view of the retina for more effective screening and grading of eye diseases. However, there are two main challenges in aligning multi-modal retinal images. First the alignment quality is influenced by the inconsistent cross-modal retinal patterns which generate ourliers in image matching, as well as the lack of anatomical boundaries which may fool the intensity-based metrics. Second, the varying distortion levels across modalities also cause misalignment that cannot be corrected by conventional global transformations, where the alignment between wide-angle and narrow-images suffers most due to a special projection model used in the wide-angle imaging process. In this thesis, we propose registration algorithms to address each of these issues. First, we use the most prominent and consistent anatomical structure, *i.e.*, retinal vasculature, as the basis for both global and deformable registration, where we train vessel segmentation networks by unsupervised style transfer techniques to achieve this purpose. Second, we propose a distortion correction module to improve global alignment quality, where wide-angle images are remapped based on the narrow-angle images' view points before registration. Furthermore, we optimize a dense 3D scene for the wide-angle images to reflect more details of the eyeball shape, such that the remapped image can be accurately aligned with the narrow-angle image.

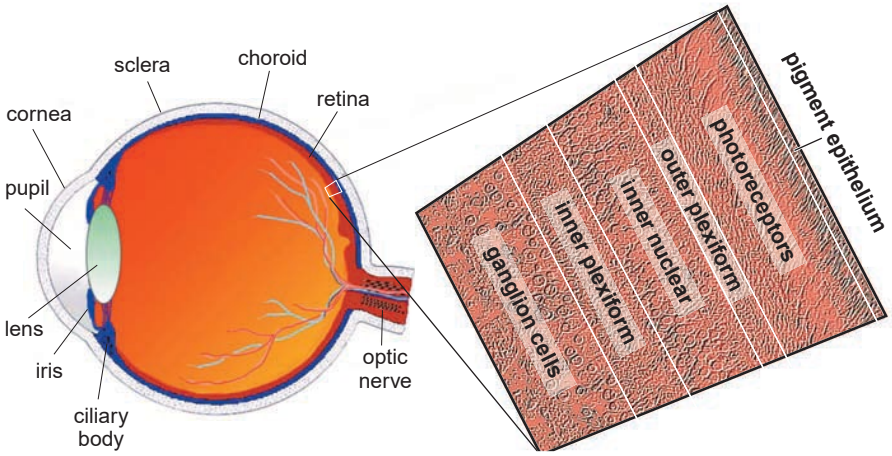


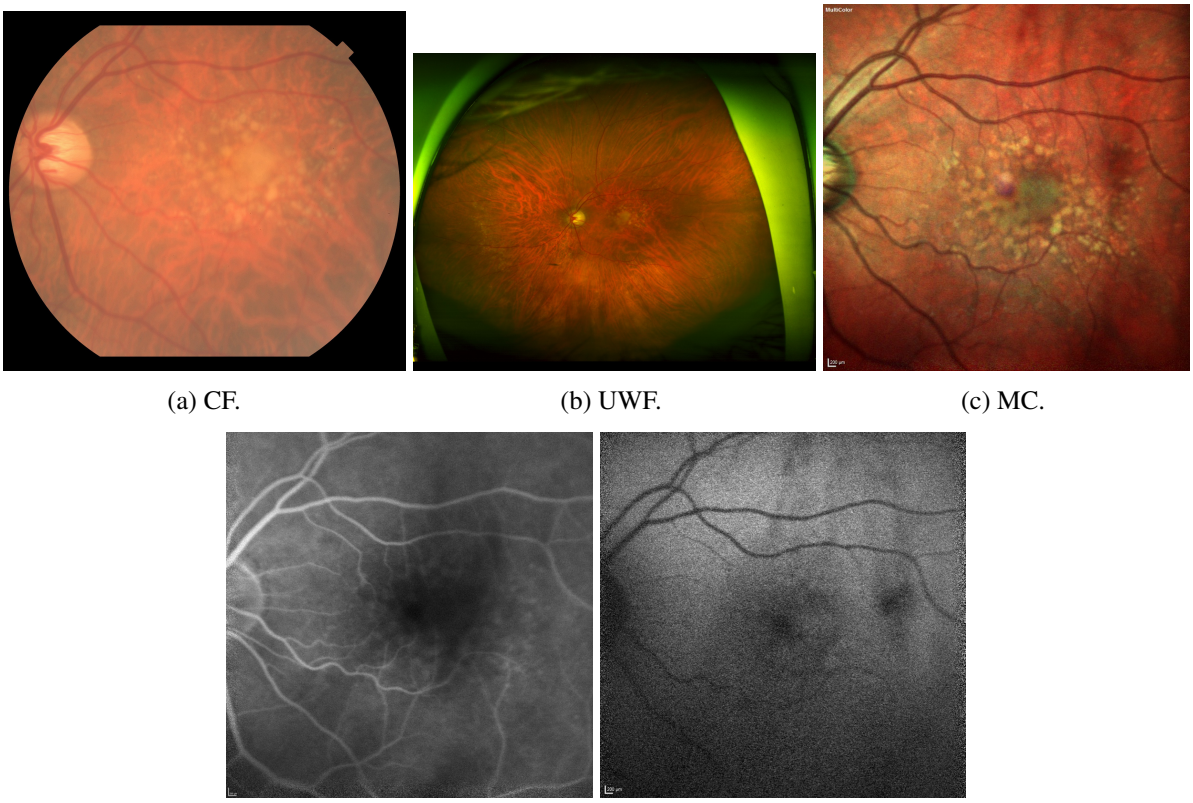
Figure 1.1: An illustration of eye structure and retinal layers. [1]

1.1 Human Eye Structure and Retinal Imaging Modalities

The human eye structure of Fig. 1.1 [1] presents that the retina is a thin layer lying on the back of the eye ball, between the vitreous body and choroid. In a more detailed partition, the retina consists of multiple layers, including the layers of photoreceptor cells (rods and cones), horizontal and bipolar cells (inner nuclear layer), ganglion cells, and nerve fibers [1]. The layer of photoreceptors is supported by the pigment epithelium and the choroid [2].

Retina plays a crucial role in human vision, as it converts lights into nerve signals which are transmitted into the brain to form visions. Since it may also suffer from various diseases that can damage human vision, including Diabetic Retinopathy (DR), Diabetic Macular Edema (DME), Age-related Macular Degeneration (AMD), Glaucoma, etc, the diagnosis and treatment of these eye diseases is important to protect or recover the visions. Specially, various retinal imaging modalities provide effective approaches for ophthalmologists to detect, recognize, and measure the diseases. Current imaging modalities for 2D retinal structures include the following aspects:

1. **Fundus camera** is a conventional and mostly-used retina imaging instrument. With the retina illuminated by white light sources, the fundus camera can capture the retina's image



(a) CF. (b) UWF. (c) MC.

(d) FA. (e) FAF.

Figure 1.2: Multi-modal images of a same eye.

in a single-shot, where reflected lights in different wavelength ranges can be captured by applying different color filters in front of the sensor.

Color Fundus (CF) modality takes an image in red, green and, blue wavelengths using the fundus camera. As the most cost-effective way for retinal imaging, CF has been widely adopted in Diabetic Retinopathy (DR) early screening [3] and Diabetic Macular Edema (DME) diagnosis. The disadvantages of CF include low contrast in retinal patterns, as well as the inability of imaging peripheral retina areas using most conventional CF cameras, as shown in sub-image Fig. 1.2 (a).

2. **Scanning Laser Ophthalmoscopy** (SLO) is an alternative imaging technique which scans and illuminates the retina by fixed-wavelength laser beams. Consequently, the captured

images suffer less noises from the optic paths and provide higher contrasts. Confocal SLO (cSLO) technique further includes a confocal aperture in front of the light receiver, which can image at certain depth in retina by suppressing out-of-focus lights [4, 5]. SLO can be used in several retinal image modalities.

MultiColor (MC) modality in the Heidelberg's Spectralis system captures pseudo color retinal images via SLO. It takes advantages of the retinal laser's varying penetration abilities at different wavelengths, *i.e.*, , lights of longer wavelengths can reach deeper retinal layers and even the choroid, while lights of shorter wavelengths are reflected on the superficial retinal layers [5]. The color channel of MC consists of three retinal images captured at the wavelengths of near infra-red, green, and blue colors, which enhances the view of retinal structures and pathologies [5]. An example of a MC image is shown in sub-image Fig. 1.2 (c).

3. **Ultra-Widefield** (UWF) imaging. Conventionally, in order to examine the peripheral retina areas, multiple narrow-angle images are captured at different positions in retina and then combined to create a wide-view image. For example, in the protocol of Early Treatment Diabetic Retinopathy Study (ETDRS) [3], separate CF images are captured at 7 pre-defined positions on retina, which increases the complexity of the operation. Alternatively, UWF imaging systems can capture most retina areas in a single shot via specially designed optics, which is faster and more convenient. Specifically, the Optos's UWF system uses an ellipsoidal mirror to create two focal points at the reflection position of laser source and human eye's pupil. When combined with the SLO technique, it can easily see the peripheral retina areas with large angles by simply adjusting the direction of the reflected laser [6]. An example of a UWF pseudo color image is illustrated in sub-image Fig. 1.2 (b).

4. **Fluorescein Angiography** (FA). By injecting fluorescein into blood vessels, the retinal vasculature can be clearly imaged by either conventional fundus cameras or SLO techniques.

As the fluorescein travels from arteries to veins through capillaries, multi-stage images can be captured to help localize eye diseases that change the vasculature. The FA modality is helpful in identifying abnormal vessels as well as leakage areas [7]. An example of a FA image is shown in sub-image Fig. 1.2 (d).

Indocyanine Green Angiography (ICGA) is an alternative modality to FA. Different from FA, ICGA uses indocyanine green dye that fluoresces in the wavelength of infra-red lights. It can help to visualize circulations in deeper retina layers, e.g. the choroidal vasculature [8].

5. **Fundus Autofluorescence** (FAF). As the retina contains several different fluorophores, autofluorescence patterns can be obtained using lights at certain wavelengths with the cSLO technique [5]. The mostly dominant fluorescence indicator in human eyes is lipofuscin in the Retinal Pigment Epithelium (RPE) which excites on blue lights and emits yellow-green lights [9]. Thus the FAF modality can help reveal diseases related to RPE and photoreceptor cells. An example of a FAF image is shown in sub-image Fig. 1.2 (e).

1.2 Multi-Modal Retinal Image Registration

Multi-modal retinal images can capture various appearance of a same anatomical structure, due to the distinct characteristics of each imaging techniques and modalities. Since aligning these image can ensemble multi-level information and create a more comprehensive view of the retina, it helps the detection of diseases and the estimation/grading of their development stages. For example, early-stage AMD causes drusen (yellow deposits) in the macular area which are visible in fundus cameras, as well as pigment changes in RPE which can be visualized in FAF modality [10]. Thus, aligning CF and FAF modalities could contribute to more accurate diagnosis and gradings of the AMD disease. In another example, CF images can provide true colors and act as the standard modality for DR screening, while UWF can only capture pseudo color images

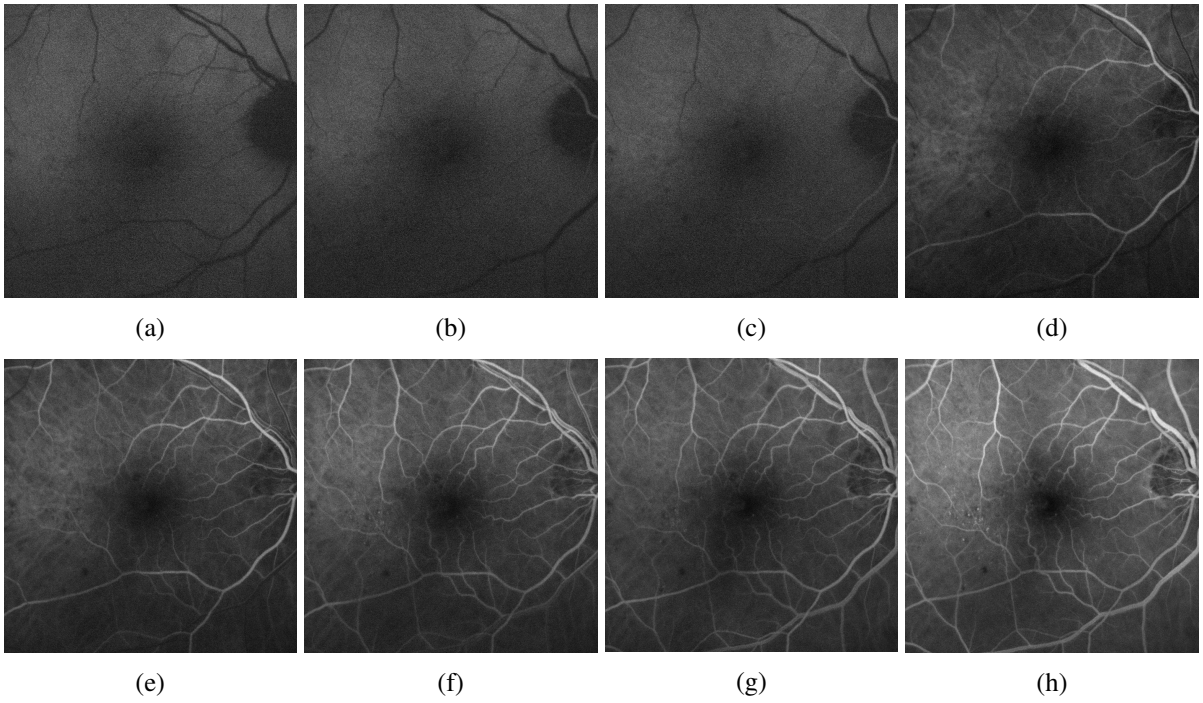


Figure 1.3: Fluorescein Angiography (FA) image sequence, captured in the order from (a) to (h).

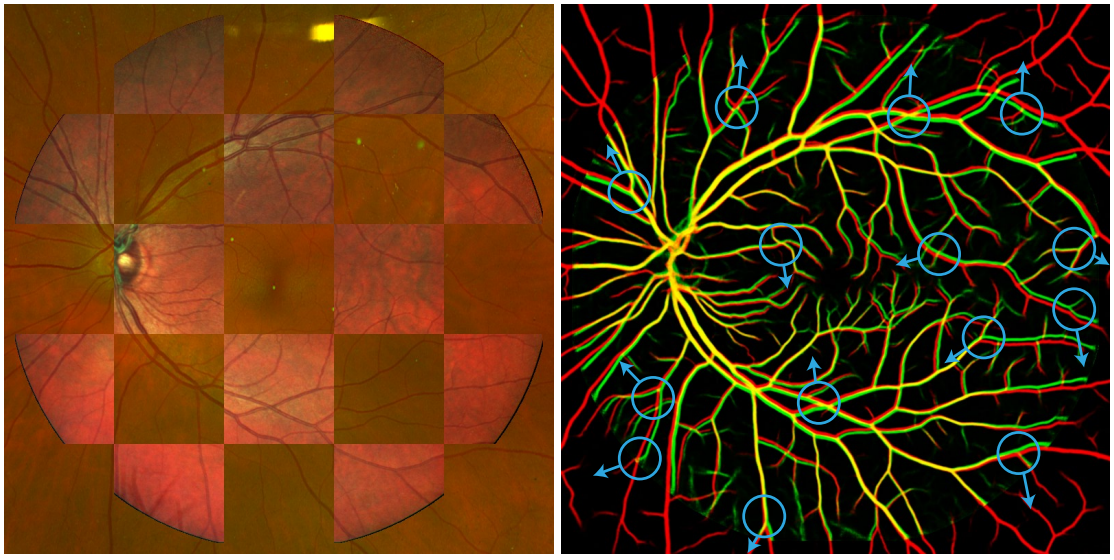
but screen larger retina areas for diseases outside of CF's coverage. The aligned CF and UWF modalities can provide both true color and wide field of view to retina, which may improve the effectiveness of DR screening.

However, there are two main challenges in multi-modal retinal image registration. First, the registration accuracies may be limited by the inconsistency in anatomical and disease patterns in various modalities, as well as the lack of object contours. As mentioned in Section 1.1, a same retinal structure may appear differently in multiple modalities. For example, (1) the retina shows low contrasts in fundus cameras (CF) but higher contrasts in SLO images (*e.g.*, UWF and MC) (Fig. 1.2). (2) Vessels have higher pixel intensities than backgrounds in FA images but lower intensities in other modalities (Fig. 1.2), as well as varying intensities at different imaging stages in the FA modality (Fig. 1.3). (3) The AMD lesions (yellow dots) in the macula area appear more prominently in the MC and CF modalities, but they are hardly visible in the FA or FAF modalities, as shown in sub-images (a), (c), (d), and (e) of Fig. 1.2. Since the retinal patterns

have varing or unmatched appearances, the keypoints and feature descriptors which are detected and extracted on these structures may become outliers in feature matching, which will undermine the registration performance.

In previous researches on multi-modal registration for medical images, multiple metrics such as Normalized Cross Correlation (NCC) and Normalized Mutual Information (NMI) [11] have been proved useful in either intensity-based optimization [12] or unsupervised deep neural network training [13, 14]. The rich anatomical patterns and the 2D boundaries (or 3D surfaces) belonging to the subjects to be aligned can function as important guidance in the intensity-based alignment process. However, those methods may fail in multi-modal retinal registration with large displacements, since there are only sparse and thin structures (vessels) and no boundaries in retinal images. Moreover, they can even be misled by the edges of the imaging area, as they will achieve local optima by aligning image contours. An example showing the weakness of intensity-based methods in aligning retina-like images is provided as Fig. 1 in [15], where the method aligns two images based on the deliberately created fake contours instead of their sparse vessel structures.

The second challenge in multi-modal retinal registration is the different levels of perspective distortions in aligning UWF images with narrow-angle images. The UWF imaging systems usually adopt a special projection process to represent the 3D surface (retina) with large curvatures onto a 2D flat array [16]. For example, the Optos UWF system uses stereographic projection to map the 3D sphere data onto a 2D plane. This mapping function can be considered as a special perspective projection in geometry, where the camera view point is set on the cornea and the objects of interest lie on the sphere (eyeball). However, this projection process also leads to large distortions in peripheral retina areas such that the peripheral patterns are expanded on the 2D plane. By contrast, other retinal modalities cover narrower view angles and set their view points more distant from the cornea, such that they contain less distortions than UWF images. This difference in distortion levels cannot be corrected by a conventional global transformation model



(a) Montaged view. (b) Overlaying of vessels.

Figure 1.4: Illustration of UWF distortions. The UWF and MC images are coarsely aligned through an affine transformation. (a) Montaged view of UWF and MC images. (b) Overlaying of the corresponding UWF (red) and MC (green) vessels, where yellow represents their overlapping areas. Blue circles and arrows indicate misalignments and the distortion's directions.

(*e.g.*, affine transformation). Besides, the feature descriptors extracted from the same retinal pattern with different distortions are more likely to be mismatched, which will also undermine the performance of feature-based alignment algorithms. An example of aligning a UWF image with a narrow-angle MC image is shown in Fig. 1.4, where the peripheral UWF vessels are over-expanded while the center UWF vessels are under-expanded.

1.3 Contributions

In this thesis, we address the two main challenges in Section 1.2. To resolve the inconsistent patterns in multi-modal retinal images, we propose to extract retinal vascular structures for both detecting features and training deformable alignment networks. On the other hand, to deal with the UWF distortions in aligning UWF and narrow-angle modalities, we propose a distortion correction module which remaps the UWF images based on the camera parameters of

the narrow-angle image. Furthermore, we propose to optimize and reconstruct a free-form 3D scene to represent the eyeball shape to further reduce the UWF image distortions.

1.3.1 Two-Step Registration on Multi-Modal Retinal Images via Deep Neural Networks

We observe that the retinal vasculature is the most outstanding indicator for multi-modal image registration. However, since most existing vessel segmentation datasets are only available in the CF modalities, the vessel segmentation networks for other modalities (such as MC and FA) cannot be trained in a fully-supervised manner. To this end, we propose a weakly supervised training scheme via style transfer [17, 18] to jointly train the vessel segmentation and deformable registration networks. The learning of the vessel segmentation task is mainly guided by a style loss [17], and the registration task is guided by the photometric consistency and smoothness loss in optical flow network training [19]

In addition, we set up a two-step coarse-to-fine registration pipeline fully based on deep neural networks. First, the coarse alignment network predicts global transformations to correct large displacements. It consists of the vesssel segmentation networks pre-trained in the fine alignment step, a pre-trained feature detection and description network [20], and an outlier rejection network [21] trained in a supervised manner. Afterwards, the deformable registration network estimates pixel-wise transformations to refine the alignment after the first step, where the network is trained in conjunction with the segmentation networks. We also investigate an alternative signal, *i.e.*, the Monogenical Phase Signals [22], as a guidance to train the deformable registration network.

1.3.2 Multimodal Global Registration between Ultra-Widefield and Narrow Angle Retinal Images via Distortion Correction Network

Since the difference in camera pose is the main cause of the peripheral distortions in UWF images when compared with Narrow-Angle (NA) images, we propose a distortion correction function for Optos UWF images. We set up pixel-wise correspondence between the original UWF image and the distortion-corrected UWF image according to stereographic as well as the more-common perspective projection. The 2D UWF pixels are first connected to 3D spherical points, which are then reprojected based on a new camera with similar extrinsics as the NA camera. We use five parameters to describe the NA camera extrinsic parameters (except one rotation around z-axis), and propose a two-stage iterative searching algorithm to find the optimal parameters.

Moreover, the distortion correction function and its searching algorithm is incorporated into the aforementioned coarse alignment network in the testing phase, which will benefit the feature detection and matching process to improve alignment performance. We also modify the deformable registration training scheme to train a vessel segmentation network for NA image when a pre-trained UWF segmentation network is available.

1.3.3 3D Eyeball Shape Estimation for Ultra-Widefield and Narrow-Angle Retinal Image Alignment

The aforementioned distortion correction function is based on the spherical assumption for the eyeball shape, while the actual eyeball shape may deviate from a pure sphere which still cause misalignments. Therefore, we set up a dense 3D mesh to represent the UWF pixels on the sphere, and update the vertices' coordinates in the mesh to find a scene which approximates the actual eyeball shape. The scene and the NA camera parameters are jointly and iteratively optimized, with the objective to reduce the error between the original UWF/NA vessel maps and

the reprojected vessel maps from the 3D scene based on UWF/NA camera poses.

Besides, we also concatenate this optimization process with the global alignment results from the previous distortion correction network. The algorithm first generates a reprojected UWF vessel map based on NA camera pose, and then warps the reprojected image by the global transformation model from the distortion correction network. With this combination, the local (dense mesh) and global (sparse parameters) alignment process can be concatenated to achieve the best performance.

1.4 Organization of the Thesis

The rest of the thesis is organized as follows. Chapter 2 presents the details of the global and deformable registration networks, including their structures and training process. It also includes a literature review on multi-modal retinal image registration methods, especially those based on deep neural networks. Chapter 3 details the UWF distortion correction module, where the mapping functions are derived and a camera parameter estimation algorithm is set up. Chapter 4 describes the 3D eyeball shape reconstruction method over UWF and NA images, which is also concatenated with the distortion correction module. Finally, Chapter 5 concludes the thesis and provides research directions to improve the current method.

2 Two-Step Registration on Multi-Modal Retinal Images via Deep Neural Networks

2.1 Introduction

Multi-modal retinal image registration plays an important role in assisting the examination and diagnosis of retina diseases. In this task, multi-modal retinal images are captured from the same patient using various retinal imaging instruments, and then aggregated and aligned, so that complementary information of the retina can be integrated for more accurate and faster inspection. In order to accurately align retinal image pairs, a two-step *coarse-to-fine* pipeline has been adopted (*e.g.*, [23, 24]) for coarse (global) alignment and fine (local) alignment, as shown in Fig. 2.1. In the coarse alignment step, a *source* (*floating*) image is warped towards a *target* (*fixed*) image based on an estimated global transformation model (*e.g.*, affine transformation). In the following fine alignment step, the globally aligned source image is warped again locally based on a pixel-wise registration field in order to further reduce misalignment errors.

A big challenge in aligning multi-modal retinal images comes from the inconsistent appearance of anatomical and diseases patterns among modalities since each instrument has different imaging mechanisms and settings. For example, in the first row of Fig. 2.2, the vessels show lower intensities than the background in the Color Fundus (CF) image, but higher intensities in the Fluorescein Angiography (FA) image. In the third example, the CF images

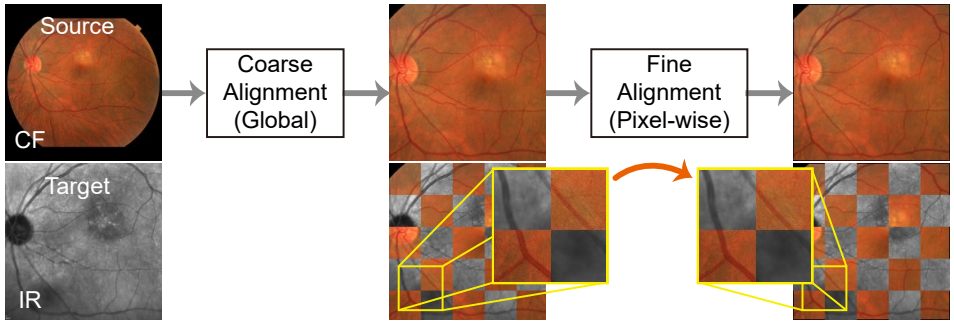


Figure 2.1: A two-step coarse-to-fine registration framework.

shows choroid patterns beneath the retinal vessels which are not visible in the Infrared Reflectance (IR) image. As a result, these inconsistent patterns will produce unmatched features and ourliers in image matching process which affects the registration performance. Furthermore, retinal images usually have thin and sparse vessels and lack dense anatomical structures, which yields multiple local maximas when computing commonly used similarities metrics such as Normalized Cross-Correlation (NCC) and Normalized Mutual Information (NMI), as shown in Fig. 2.2. This also causes difficulties for intensity-based registration methods in finding the correct alignment.

There has been extensive research on multi-modal retinal image registration. For example, many methods have used hand-designed algorithms to replace certain steps in a conventional registration pipeline, including keypoint detection (*e.g.*, UR-SIFT [25]), hand-crafted feature description (*e.g.*, PIIFD [26], and Step Patterns [27]), and matching and outlier rejection (*e.g.*, [28, 29]). Meanwhile, some have also tried to utilize the mutual structures in paired images to aid registration, including vessels [24, 30–32], and vessel bifurcations and crossovers [24]. Nevertheless, many of these methods lack robustness in challenging scenarios like poor imaging qualities.

Recently, with the rapid development of deep learning techniques, some methods have applied Deep Neural Networks (DNNs) in this task, as summarized in Table 2.1. However, the restrictions of each method limit its application on general multi-modal retinal datasets as follows: **(1)** The authors of [34–36] applied Convolutional Neural Networks (CNNs) for parts

Table 2.1: Comparison of Multi-Modal Retinal Registration Algorithms Adopting Deep Learning

Method	Transformation Model	Working Modalities	Required Annotations / Inputs	Fully Network	Method Description	Major Limitations
Mahapatra *et al.* [33]	Deformable	No restriction	Accurately aligned images	Yes	GAN for warped image synthesis	No explicit warping process
Lee *et al.* [34]	Global (affine)	No restriction	Not required	No	CNN for feature selection in a conventional pipeline	Performance limited by conventional methods
Arikan *et al.* [35]	Global (affine)	No restriction	Vessel segmentations + keypoint locations	No	CNN for vessel segmentation and keypoint detection, plus RANSAC	Demanding massive labeling
Ding *et al.* [36]	Global (polynomial)	UWF CF & FA	Pre-trained vessel segmentation CNN for FA	No	CNN for vessel segmentation, plus a conventional alignment method	Requiring a pre-trained segmentation network
Luo *et al.* [37]	Global (affine)	ICGA + MC	Affine matrix + optic disc segmentation	Yes	CNN for optic disc segmentation and matrix estimation	Initialized by camera-specific scaling factors
Tian *et al.* [38]	Deformable	No restriction	Coarsely aligned images	Yes	CNN for optical flow estimation	Only for small displacements
Zhang *et al.* [39]	Deformable	No restriction	Affine matrix or coarsely aligned images	Yes	CNN for vessel segmentation and optical flow estimation	Only for small displacements
Wang *et al.* [40]	Global (perspective)	No restriction	Affine matrix + pre-trained vessel segmentation CNN	Yes	CNN for vessel segmentation, feature detection & description, and outlier rejection	Requiring pre-trained segmentation networks
Ours	Global (affine) + Deformable	No restriction	Affine matrix + a style target image	Yes	CNN for vessel segmentation, feature detection and description, outlier rejection, and optical flow estimation	

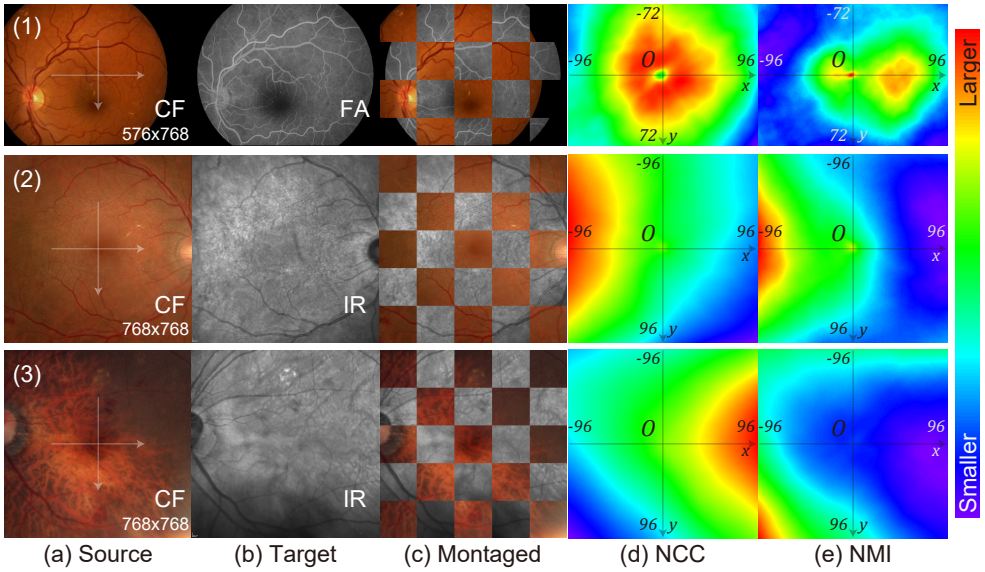


Figure 2.2: Similarity measurements on multi-modal retinal images with regard to image translation. First, we coarsely align the source images (a) with the target images (b), which are overlaid as (c). Then, we translate the source images in both x and y directions by different number of pixels. At each position, we compute Normalized Cross-Correlation (NCC) (d) and Normalized Mutual Information (NMI) (e) between the two images, and plot their heatmaps. We only use the overlapping pixels in the imaging area of both images to compute the similarity metrics. First row is from the CF-FA dataset. Second and third rows are from the JRC dataset. In these examples, both NCC and NMI should be the highest at the center (0,0) to correctly estimate alignment.

or additional modules of a conventional global registration pipeline, whose performance is still limited by the conventional algorithms. **(2)** Many methods require massive labeling works (*e.g.*, pixel-wise alignment or segmentation) for training which is hard to achieve in large-scale datasets, *e.g.*, [35, 37] demand explicit segmentation ground-truths, [36, 40] adopt pre-trained vessel segmentation networks which implicitly demand segmentation labels, and [33] requires accurately aligned image pairs. Only a few methods [38, 39] need coarsely aligned images which are easy to obtain (*i.e.*, requiring affine matrices by labeling the positions of three corresponding point pairs). **(3)** Some methods [36, 37] contain camera-specific designs which might restrict their application in general cases. **(4)** All the previously proposed methods only handle one step (coarse only or refinement only), *i.e.*, [34–37, 40] only support the global transformation which limits their registration accuracies, and [33, 38, 39] tackle deformable registration with

unsupervised training which would fail on images with large displacements. To our knowledge, there are no DNN methods to support both global and deformable alignment.

To this end, we propose a two-step coarse-to-fine registration algorithm for multi-modal retinal images as shown in Fig. 2.1. The proposed method only requires easy-to-obtain annotations (*i.e.*, affine matrix) for training, and is completely built upon DNN which eliminates restrictions of the conventional methods. In the coarse alignment step, vessels of source and target images are first extracted via vessel segmentation networks. Then, keypoints and features of the vessels are derived using a feature detection and description network. Afterwards, a transformation matrix is estimated by an outlier rejection network where the matrix is used to warp the source image to the target image for the global alignment. In the next fine alignment step, a deformable registration network predicts a pixel-wise registration field to warp the source image for a second time to further reduce misalignment errors. In the learning process, we also propose a high-level structure, which is named modality transformer, to handle different appearance of modalities and the lack of pixel-wise ground-truths. The modality transformers can find common structures between multi-modal images to enable unsupervised training for the deformable registration network. We set up two kinds of transformers in this chapter, *i.e.*, a non-learnable local phase signal extractor, as well as the vessel segmentation networks which are trained jointly with the deformable network using style loss [17, 18]. Meanwhile, the feature detection and description network is trained on a large-scale synthesized dataset, and the outlier rejection network is trained with ground-truth transformation matrices on the retinal datasets.

In this chapter, we expand our previous works [39, 40] to support both coarse and refinement registration in the following aspects. (1) For the deformable alignment task [39], we set up an unsupervised learning framework consisting of the deformable registration network and two modality transformers. During training, the modality transformers convert multi-modal images into image signals of a common modality to compute photometric consistency loss. (2) We also propose a Local Phase Modality Transformer that extracts local phase terms in Monogenical

Signals [22] for the deformable registration task. (3) We enhance the network structures and training settings from [39], and extend the experiments to show the influences on deformable registration performance from various factors, including smoothness weights, number of local phase's channels, and the choice of style targets. (4) We also include extensive ablation studies for the global registration network comparing to our previous publication [40]. (5) We combine the coarse alignment [40] and the deformable registration [39] methods into a complete pipeline, and train and test it on a newly collected multi-modal retinal dataset.

The chapter is organized as follows. Section 2.2 introduces backgrounds and related works for multi-modal retinal image registration. Sections 2.3 and 2.4 describe algorithmic details of the proposed coarse and fine alignment networks. Section 2.5 presents experimental results and ablation studies on our methods.

2.2 Backgrounds and Related Works

2.2.1 2D Image Registration

2D image registration algorithms can be categorized into feature-based and intensity-based methods. For feature-based methods, most conventional algorithms follow a fixed non-iterative routine which consists of keypoint detection (*e.g.*, Harris corner detector [41]), feature description (*e.g.*, SIFT [42]), feature matching, and outlier rejection (*e.g.*, RANSAC [43]). In intensity-based methods, a correlation metric (*e.g.*, Mutual Information) is designated to evaluate the alignment quality between an image pair, and an iterative optimization algorithm helps search for a set of warping parameters that achieves the best quality of alignment. Usually, the latter approaches are much more time-consuming, because the searching process could not be done in parallel.

There are mainly two types of transformation models that describe the warping process on the source image: global and deformable. In global transformation, the movement of all pixels are determined by a set of global parameters such as scaling, translating, rotation, and skewing of

the affine transformation. In deformable transformation, the pixel-wise registration fields (optical flows) are estimated, and each pixel of a source image is warped to the target pixel using its own optical flow that describes the direction and the distance.

In this chapter, the proposed two-step framework adopts an affine transformation for coarse alignment and a deformable transformation for fine alignment. Both steps are accomplished through non-iterative processes.

2.2.2 Global Registration for Natural Images

Recently, much effort has been made in adapting deep neural networks for the global registration tasks. Most methods comply with the feature-based registration pipeline explicitly or implicitly. Some have trained networks to replace certain steps in the registration pipeline, *e.g.*, descriptors [44, 45], outlier rejection [46, 47], descriptors with matching metrics [48], detectors with descriptors [20, 21, 49, 50]. Moreover, other methods (*e.g.*, CNN-Geometric [51, 52]) proposed to replace the complete pipeline with an end-to-end network. Nevertheless, in order to achieve good registration results on cross-modality tasks, these methods require large-scale labeled data for training, or need pre-trained network-based descriptors which can extract robust features from multi-modal images.

In our proposed coarse alignment method, SuperPoint [20] is adopted as the keypoint detector and descriptor, and the outlier rejection network [47] is trained to estimate the transformation matrix. Especially, two vessel segmentation networks help translate multi-modal retinal images into single-modal vessel images as SuperPoint's input, so that SuperPoint only needs training on synthesized single-modal data instead of labeled multi-modal data.

2.2.3 Optical Flow Estimation for Deformable Registration

Optical flow estimation computes a dense registration field between the source and target images of a same modality. It is built on the assumption of brightness consistency between the

two images. Conventional algorithms (*e.g.*, [53]) often involve an iterative searching process over a loss function which optimizes photometric consistency and smoothness constraints in deformable alignment.

Recently, multiple CNN-based methods have been proposed to learn a set of parameters on training data and to eliminate the iterative optimization process during testing. The network can be trained by a supervised scheme or an unsupervised scheme. Some methods [54–56] adopt the supervised training on large-scale synthesized datasets with ground-truth flows, which enables them to handle larger displacements. Meanwhile, others [19, 57] adopt the photometric consistency and smoothness loss for the unsupervised training (*i.e.*, without ground-truth labels), which are limited to predict small displacements. Spatial Transformer Networks (STN) [58] is often used as a differentiable image warper in the unsupervised learning scheme.

In this chapter, we adopt the unsupervised training method for the fine alignment step of our framework with the help of modality transformers (*i.e.*, vessel segmentation networks or local phase signals).

2.2.4 Medical Image Registration

In contrast to aligning natural images, the intensity-based techniques form the vast majority of conventional registration methods on medical images [12]. Widely-used similarity metrics include NCC, Mutual Information (MI), and NMI [11], etc, which can be applied to both mono- and multi-modal registration.

Recently, multiple CNN-based methods [13, 14, 59–62] have also been proposed for medical image registration in one shot, *i.e.*, non-iteratively. These methods adapts the unsupervised learning scheme in Section 2.2.3 by replacing the photometric consistency loss with other aforementioned similarity metrics. Furthermore, anatomical structures within the images can also be extracted and compared during training [60] to boost performance. Among these methods, the networks proposed in [13, 59, 60] only perform one-time registration, which are limited to

predicting small displacements. Instead of using only one network, de Vos *et al.* [14] proposed a coarse-to-fine registration framework which concatenates an affine registration network and multiple deformable networks. Zhao *et al.* [61, 62] also proposed a recursive cascaded network where the floating image is warped progressively by multiple cascades, which enables predictions for large displacements.

It should be noted that, due to the nature of the similarity metrics, these methods achieved success by correlating the dense anatomical structures between images. In addition, when aligning subjects with large displacements, they rely on the subjects' surfaces/contours to find the initial warping direction. However, they are not suitable for retinal images which have sparse and thin vessels and lack subject boundaries (*e.g.*, [63]), because they will be trapped at local maximas in the searching space of similarity measurements during optimization. Fig. 2.2 shows simple examples of NCC and NMI measurements when aligning retinal images by 2D translation. In the first row, there are two local maxima areas in the NMI heatmap, in which the right one corresponds to the wrong alignment based on the imaging circles. In the second and thrid rows, the local maximas in the center (*i.e.*, the correct alignment position) become much less obvious (the second row) or even invisible (the third row), which will mislead the algorithms into wrong warping directions. Lee *et al.* [15] also shared a similar observation in their work.

2.2.5 Multi-Modal Retinal Image Registration

Table 2.1 summarizes multi-modal registration methods based on deep learning, and none of them addresses the complete coarse-to-fine registration pipeline. Lee *et al.* [34] proposed a feature filtering CNN to detect and remove unreliable step features for multi-modal retinal image registration. Specifically, their network is trained with image patches as inputs and their corresponding step patterns [27] as outputs. During testing, unreliable patches are removed if their predicted patterns from the network deviate from any possible patterns. Arikan *et al.* [35] proposed to align multi-modal retinal images based on vessel segmentations and bifurcations

using two CNNs. However, the networks are trained with segmentation and bifurcation ground-truths which are difficult to obtain in most cases. Luo *et al.* [37] proposed a CNN to estimate affine transformation matrix for IndoCyanine Green Angiography (ICGA) and Multi-Color (MC) images. However, it requires optic disc segmentations and dataset-specific scaling parameters for training, which limits its application on other datasets. Tian *et al.* [38] proposed a deformable retinal registration network, which adopts image gradients of two images as alignment signals for unsupervised training. Mahapatra *et al.* [33] proposed a deformable registration model based on unsupervised CycleGAN [64]. Instead of predicting a registration flow field, their network directly synthesizes a warped floating image, which cannot be used for diagnose purpose since the results might include non-existing patterns and lose critical lesion diseases. They also attached an additional branch to the network to predict registration fields which requires accurately aligned images for training, which is not applicable in most cases.

Particularly, Ding *et al.* [36] proposed to train a vessel segmentation network for Ultra-Wide Field (UWF) CF images through a joint segmentation and registration scheme, which bears similarities to the fine alignment network of our previous publication [39] and the second step of the proposed framework in section 2.4.2. In brief, with paired UWF CF and FA images as well as a pre-trained vessel segmentation network for FA, the vessel ground-truths for CF are obtained from the vessel predictions on FA images by aligning FA vessels with CF vessel predictions. The vessel segmentation network for CF and the alignment process are trained and optimized iteratively. However, their method mainly focuses on vessel segmentation instead of multi-modal registration, and is designed for UWF images from a same optic system which have similar scales and resolutions. It relies on a good initialization (*i.e.*, a pre-trained segmentation network for one modality) which is hard to obtain in general cases. Besides, it adopts a global transformation model (with 12 parameters) which is insufficient for images bearing larger distortions (*e.g.*, images from different instruments). In comparison, our proposed method does not require good initialization for segmentation and adopts a two-step coarse-to-fine structure, which is a more

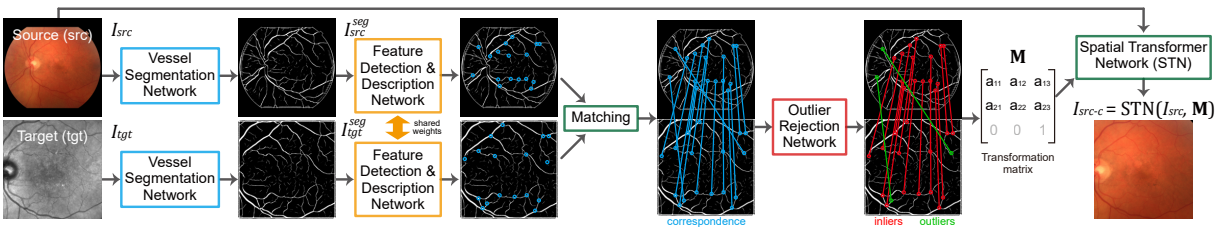


Figure 2.3: The coarse alignment step of the proposed two-step framework.

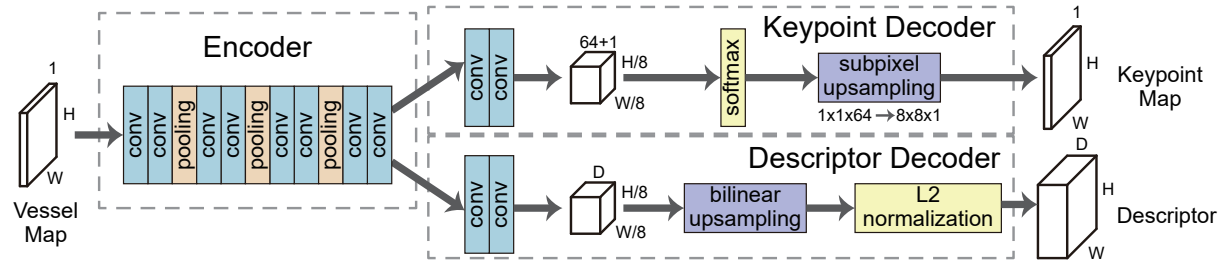
flexible and general solution for multi-modal retinal registration.

2.3 Two-Step Framework: Coarse Alignment

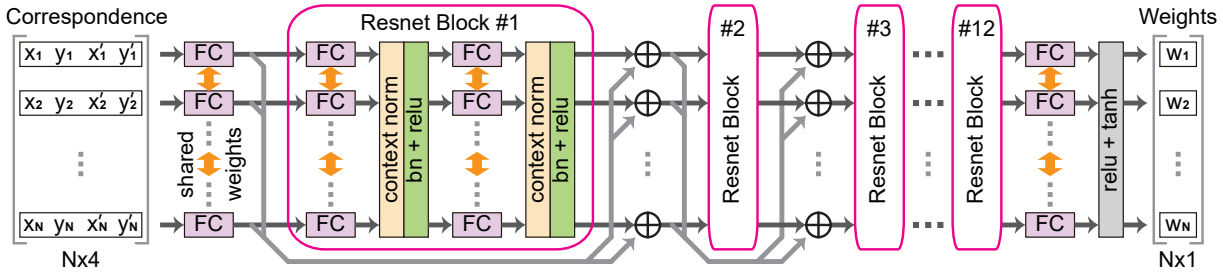
The proposed coarse alignment algorithm consists of three sequentially concatenated networks for vessel segmentation, feature detection and description, and outlier rejection as is shown in Fig. 2.3. First, the vessel segmentation network transforms multi-modal images into a common modality (*i.e.*, grayscale vessel maps). Next, the feature detection and description network finds sparse keypoints from the vessel maps and extracts features on all keypoints. Then, features from source and target images are matched against each other based on their similarities. Finally, the outlier rejection network finds the correct matches (inliers) and removes the incorrect ones (outliers) such that an accurate affine matrix can be estimated from the inliers.

2.3.1 Vessel Segmentation

Vessel extraction has become the basis of multiple retinal registration algorithms [24, 30–32, 35], because vessels are usually the most prominent and useful patterns in multi-modal retinal images. Even though DNN has achieved great performance on retinal vessel segmentation with supervised training, its performance is not guaranteed on test data of other modalities unseen during training. Besides, we can hardly find any segmentation datasets on modalities other than CF, while it is labor-intensive to label the vessel segmentation for new datasets. Therefore, we propose an unsupervised scheme to train two vessel segmentation networks for each input



(a) Feature detection and description network (SuperPoint [20])



(b) Outlier rejection network [47]

Figure 2.4: The feature detection and description network and the outlier rejection network of the coarse alignment step.

modality without segmentation ground-truths. Specifically, the segmentation networks are trained jointly with a deformable registration network in Section 2.4. Style loss [17, 18] is used as a guidance for the segmentation task. Details on network structures and loss functions are presented in Section 2.4.2.

2.3.2 Feature Detection and Description

SuperPoint [20] network is adopted as our feature detector and descriptor, whose structure is shown in Fig. 2.4 (a). The network consists of an encoder which takes the grayscale vessel map from the previous segmentation network, and two decoders which predict a keypoint probability map and a descriptor tensor respectively. Accordingly, two loss functions are designed to train the network, including a keypoint loss and a descriptor loss. The keypoint loss penalizes missed or wrong keypoint predictions through a cross-entropy loss. Meanwhile, the descriptor loss maximizes the similarities between features of matching points or vice versa through a hinge loss. Readers could refer to [20] for more details.

Since ground-truth keypoints for our retinal images are not available, we directly use a SuperPoint model which is pre-trained on a large-scale synthesized dataset [20]. Specifically, the training dataset consists of rendered images with grayscale shapes and their ground-truth corners, which bear much similarity with our vessel maps. Therefore, the trained model can be directly applied to extract keypoints and features from the vessel segmentation results. As a post-processing step, non-maximum suppression thresholded at 5 pixels is applied over the keypoint probability maps, and pixels with confidence larger than 0.015 are denoted as keypoints. Afterwards, the corresponding feature vectors of the keypoints from both images are matched against each other based on minimum euclidean distances through a bi-directional search, e.g. feature A from the source should be the best match for feature B from the target, and vice versa. Finally, the corresponding coordinates of the matched keypoint pairs are forwarded to the next outlier rejection network.

2.3.3 Outlier Rejection Network

To obtain an accurate transformation matrix, we adopt and train an outlier rejection network [47] to detect and eliminate outliers from the matching pairs. First, the network takes the coordinates of the matched keypoint pairs from the feature detection and description network, and outputs their probabilities of being inliers. Then, the affine matrix is computed from the weighted coordinates. During training, in addition to the classification loss on the predicted weights and the regression loss on the estimated matrix, we also propose a Dice loss which evaluates the alignment quality based on the estimated matrix.

The structure of the network is shown in Fig. 2.4 (b). The network has 12 consecutive residual blocks. In each block, there are 2 fully connected layers with shared weights among N different entries, where N is the number of correspondences. Each fully connect layer is followed by a context normalization layer [47] and a batch normalization layer. Specifically, the weight-sharing design ensures that each correspondence will be processed independent of its

input order. Meanwhile, the context normalization layers enable the sharing of global context among all correspondences.

In the forward process, the network takes a matrix $[\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_N]^T \in \mathbb{R}^{N\times 4}$ as input, where $\mathbf{q}_i = [\mathbf{p}_i^T, \mathbf{p}_i'^T]^T$, and $\mathbf{p}_i = [x_i, y_i]^T$ and $\mathbf{p}_i' = [x_i', y_i']^T$ are the source and target keypoint coordinates respectively for the i-th correspondence. The network's output is a vector $[o_1, o_2, ..., o_N]^T \in \mathbb{R}^{N\times 1}$, which is further translated into a weight vector $\mathbf{w} = [w_1, w_2, ..., w_N]^T$, with each element $w_i = \tanh\big(\text{ReLU}(o_i)\big) \in [0, 1)$ being a weight for its input correspondence. Larger weights indicate more importance in estimating the affine matrix, and zero weights indicate outliers. Afterwards, an affine matrix $\mathbf{M} \in \mathbb{R}^{2\times 3}$ can be solved via weighted least square method based on the correspondences' coordinates and their weights $\mathbf{w}$, *i.e.*, solving

$$\arg\min_{\mathbf{M}} \big(\mathbf{b} - \mathbf{A}\text{Vec}(\mathbf{M})\big)^T \mathbf{W} \big(\mathbf{b} - \mathbf{A}\text{Vec}(\mathbf{M})\big) \tag{2.1}$$

where $\text{Vec}(\mathbf{M})$ is the vectorized $\mathbf{M}$, $\mathbf{b} = [x_1', y_1', ..., x_N', y_N']^T$ $\in \mathbb{R}^{2N\times 1}$, $\mathbf{A} \in \mathbb{R}^{2N\times 6}$ is constructed as

$$\begin{pmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_1 & y_1 & 1 \\ ... & & & & & ... \\ x_N & y_N & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_N & y_N & 1 \end{pmatrix}, \tag{2.2}$$

and $\mathbf{W} = \text{diag}([w_1, w_1, ..., w_N, w_N]) \in \mathbb{R}^{2N\times 2N}$ is a diagonal matrix. The solution to Eq. (2.1) is

$$\text{Vec}(\mathbf{M}) = \big(\mathbf{A}^T\mathbf{W}\mathbf{A}\big)^{-1}\big(\mathbf{A}^T\mathbf{W}\mathbf{b}\big). \tag{2.3}$$

It should be noted that we adopt the simpler affine transformation instead of the perspective

transformation in [40], because the misalignment errors can be corrected by the following fine alignment step.

In order to train the network, three loss functions are combined, including a classification loss, a regression loss and a Dice loss. The classification loss is defined as

$$L_c = \frac{1}{N}\sum_{i=1}^{N} \gamma_i \mathrm{BCE}\big(y_i, \sigma(o_i)\big) \tag{2.4}$$

where $\mathrm{BCE}(\cdot)$ is binary cross entropy loss, $\sigma(\cdot)$ is sigmoid function, $y_i \in \{0,1\}$ is the inlier ground-truth, and γ_i is a weight to balance positive and negative samples. The inlier ground-truth is obtained based on the ground-truth affine matrix $\mathbf{M}_{gt}$ as

$$y_i = \begin{cases} 1, & ||T(\mathbf{p}_i, \mathbf{M}_{gt}) - \mathbf{p}'_i|| \leq 5\,\text{pixels} \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

where $T(\mathbf{p}_i, \mathbf{M}_{gt})$ calculates the corresponding coordinate in target image for the source point $\mathbf{p}_i$ based on $\mathbf{M}_{gt}$, *i.e.*, a keypoint pair with distance no more than 5 pixels in the target image after warping are denoted as inliers.

In addition to the loss on the predicted weights, the regression loss penalizes the mean squared error (MSE) of the estimated affine matrix $\mathbf{M}$ from the ground-truth $\mathbf{M}_{gt}$, which is defined as

$$L_r = \mathrm{MSE}(\mathbf{M}_{gt} - \mathbf{M}). \tag{2.6}$$

Moreover, a Dice loss is proposed to check the alignment quality of the target and source vessel map after warping based on the estimated matrix. The Dice coefficient is defined as

$$\mathrm{DICE}(I_1, I_2) = \frac{2\left|I_1 \bigcap I_2\right|}{\left|I_1\right| + \left|I_2\right|}, \tag{2.7}$$

where I_1 and I_2 must be binary images in this function. Since our vessel maps are grayscale

images, we define a Soft Dice function by relieving the binary constraint over I_1 and I_2 as

$$\text{DICE}_s(I_1, I_2) = \frac{2 \cdot \sum \big(\text{ele\_min}(I_1, I_2)\big)}{\sum I_1 + \sum I_2}, \tag{2.8}$$

where ele_min$(\cdot,\cdot)$ takes the element-wise minimum values across the two images, and I_1, I_2 are vessel probability maps. The Dice loss is defined as

$$L_D = 1 - \text{DICE}_s\big(\text{STN}(I_{src}^{seg}, \mathbf{M}), I_{tgt}^{seg}\big) \tag{2.9}$$

where STN$(\cdot,\cdot)$ is the non-parametric differentiable image warping function [58], I_{src}^{seg} and I_{tgt}^{seg} are vessel segmentation maps of source and target images respectively, as denoted in Fig. 2.3. Finally, the total loss is written as

$$L = \lambda_c L_c + \lambda_r L_r + \lambda_D L_D \tag{2.10}$$

where λ_c, λ_r and λ_D are weighting factors.

2.4 Two-Step Framework: Fine Alignment

Due to lack of accuracy in the estimated matrices, image distortion from imaging instruments, and various field of view, there are still registration errors between the warped source image and the target image after the coarse alignment step. Many of these errors are minor and exist in local areas, which are hard to be corrected by global transformation. Since a fine alignment step using deformable transformation is necessary to further reduce these misalignment, we propose an unsupervised learning framework to train a deformable registration network and introduce modality transformers to aid the training process for multi-modal retinal images.

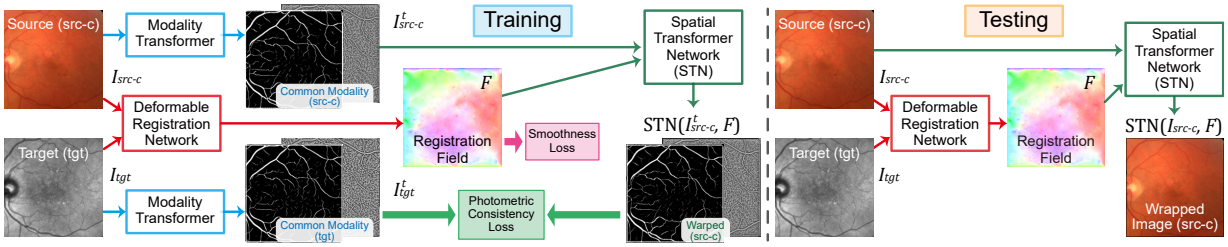


Figure 2.5: Training and testing phase for fine alignment framework.

2.4.1 Unsupervised Learning Framework

The proposed learning framework of Fig. 2.5 for training consists of a deformable registration network and two modality transformers for the source image $I_{src\text{-}c} = \text{STN}(I_{src}, \mathbf{M})$ and the target image I_{tgt}. The deformable registration network takes the multi-modal retinal image pairs as input, and predicts a pixel-wise registration field F. This is similar to optical flow networks [19] except that two input images fail to meet the brightness consistency assumption for optical flow estimation. Therefore, the photometric consistency loss cannot be directly applied on the inputs for unsupervised training. However, the proposed modality transformers change muti-modal inputs into common modality images, $I^t_{src\text{-}c}$ and I^t_{tgt}, $t \in \{seg, phase\}$, which can maintain pixel-wise correspondence as illustrated in Fig. 2.5. This helps to satisfy the brightness consistency constraint on input images like the optical flow estimation since the photometric consistency loss needs to be evaluated over their transformed modality during training. In this chapter, two different modality transformers are proposed, *i.e.*, vessel segmentation networks in Section 2.4.2, and Monogenical Phase Signal extractors in Section 2.4.3. The training process for the registration network does not require ground-truth flows for supervision, which is similar to the methods in [19, 57].

Two loss functions are used to train the registration network, *i.e.*, photometric consistency loss and smoothness loss. The photometric consistency loss is defined as

$$L_{pc}(I^t_{src\text{-}c}, I^t_{tgt}, F) = \text{MSE}\big(\text{STN}(I^t_{src\text{-}c}, F), I^t_{tgt}\big). \tag{2.11}$$

In brief, it takes the difference between the common structures extracted from the warped source image and the target image as the supervision for training. Meanwhile, the smoothness loss is defined as

$$L_{sm}(F) = \underset{k,i,j}{\text{mean}}\left((F_{k,i,j} - F_{k,i+1,j})^2\right) + \underset{k,i,j}{\text{mean}}\left((F_{k,i,j} - F_{k,i,j+1})^2\right) \tag{2.12}$$

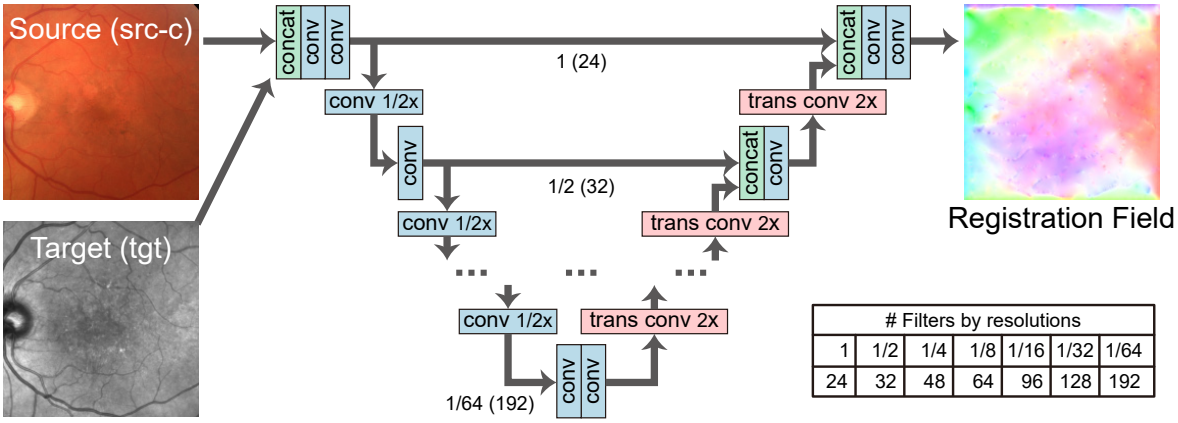
where F has dimension $2 \times h \times w$, and k, i, j are indices in F. It forces neighboring pixels in an estimated registration field to share similar warping directions and magnitudes. Therefore, displacement vectors for areas lacking details (*e.g.*, non-vessel areas in a vessel segmentation map) can be estimated from their neighboring areas (*e.g.*, areas containing vessels). In the case of using non-learnable modality transformers (*e.g.*, local phase signals), the total loss of the deformable registration network is written as

$$L_{Def} = \lambda_{pc} L_{pc} + \lambda_{sm} L_{sm}, \tag{2.13}$$

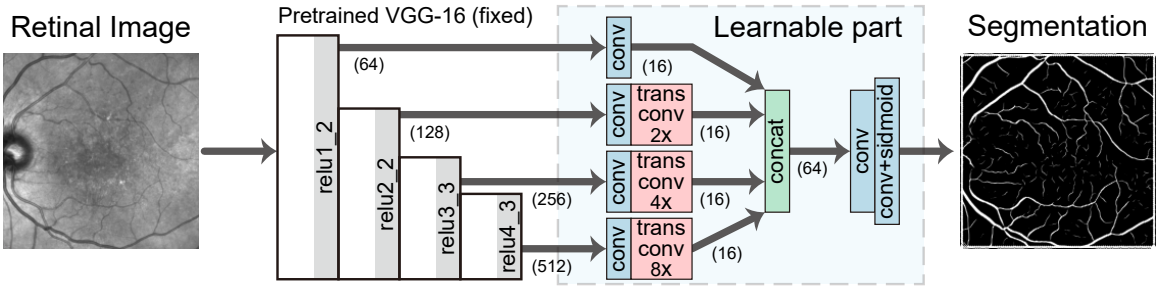
where λ_{pc} and λ_{sm} are weighting factors. In this chapter, we adopt a modified U-Net [65] as our fine registration network as illustrated in Fig. 2.6 (a). In brief, the network first extracts multi-scale features from a concatenation of two multi-modal images, where convolutional layers with stride 2 are used as downsampling layers. Then, it gradually upsamples the features at each scale and concatenates them with features from a higher scale, where transposed convolutional layers are used for upsampling. Finally, it estimates the registration field F which has the same spatial size as the input images.

2.4.2 Modality Transformer: Vessel Segmentation Network

In this section, we adopt segmentation networks as the modality transformers to guide registration and propose an unsupervised learning scheme, which is based on style transfer [17,18], to train the segmentation networks jointly with the registration network without segmentation



(a) Deformable registration network



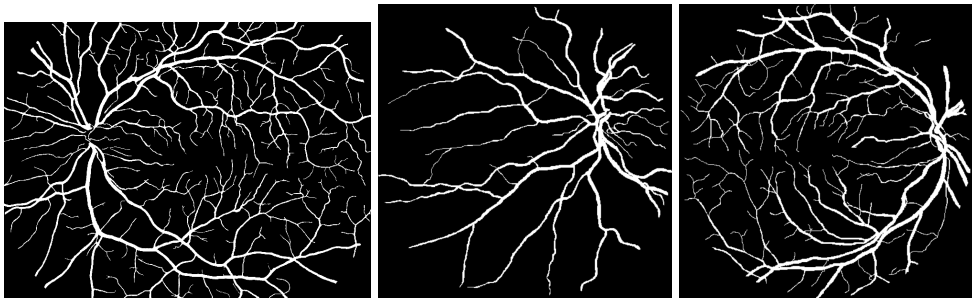
(b) Modality transformer: Vessel segmentation network

Figure 2.6: Network structures for fine alignment.

ground-truths. Specifically, the segmentation network is trained through a style loss [18] which penalizes the style difference between the network output and a style target. First, a pre-trained VGG-16 [66] network ϕ takes an image I and computes a feature tensor from its j-th layer as $\phi_j(I)$ with shape $c_j \times h_j \times w_j$. Then, $\phi_j(I)$ is reshaped into a matrix $\phi_j(I)$ with shape $c_j \times (h_j w_j)$. Next, the style feature of I is represented by a $c_j \times c_j$ Gram matrix $\mathbf{G}_j(I)$ as

$$\mathbf{G}_j(I) = \frac{1}{c_j h_j w_j} \phi_j(I) \phi_j^T(I) \tag{2.14}$$

where the spatial information in $\phi_j(I)$ is removed and only the information on global style distributions (*e.g.*, vessel-like structures) are preserved. Finally, the style loss is derived by



(a) HRF-12h [67] (b) DRIVE-23 [68] (c) DRIVE-28 [68]

Figure 2.7: Style target images I_{style} taken from publicly available datasets.

minimizing the difference of two style distributions as

$$L_{sty}^{j}(I_1, I_2) = ||\mathbf{G}_j(I_1) - \mathbf{G}_j(I_2)||_F^2 \tag{2.15}$$

where $||\cdot||_F$ computes Frobenius norm over a matrix. In our case, I_1 is a segmentation network's prediction ($I_{src\text{-}c}^{seg}$ or I_{tgt}^{seg}), and I_2 is a style target I_{style}, *i.e.*, one of the vessel images in Fig. 2.7. Therefore, the segmentation network should produce an output which also demonstrates vessel-like appearance but no pixel-wise correspondence with the style target.

In addition to the style loss, we propose a self-comparison loss to enforce rotation invariance on the vessel segmentation results. Since image edge filters show directional dependency, it is necessary to enforce the learning of edge filter pairs with inverse directions when training without ground-truths such that both edges of the vessels can be extracted. Therefore, the self-comparison loss is defined as

$$L_{com}(I) = \mathrm{MSE}\Big(\mathrm{rot}\Big(\mathrm{H}\big(\mathrm{rot}(I)\big)\Big), \mathrm{H}\big(I\big)\Big) \tag{2.16}$$

where $\mathrm{H}(\cdot)$ is the segmentation network, and $\mathrm{rot}(I)$ rotates the input image by 180°. When we jointly train the segmentation networks and the registration network, we include the style loss L_{sty} and the self-comparison loss L_{com} to the total loss L_{Def}. As a result, the total loss function is

defined as

$$L_{Def\text{-}Seg} = L_{Def} + \lambda_{com} \sum_{x} L_{com}(I_x) + \lambda_{sty} \sum_{x,j} L^{j}_{sty}(I^{seg}_{x}, I_{style}) \tag{2.17}$$

where $j \in \{\text{relu1\_2}, \text{relu2\_2}, \text{relu3\_3}, \text{relu4\_3}\}$ are VGG-16 layers and $x \in \{src, tgt\}$. Through the joint training, the segmentation networks can achieve better performance, as the segmentation prediction $I^{seg}_{src\text{-}c}$ is supervised by both the style constraints L^{j}_{sty} and another segmentation map I^{seg}_{tgt}, and vice versa. We adopt a modified network structure based on DRIU [69] for the segmentation network as shown in Fig. 2.6 (b). The network first extracts multi-scale features using a pre-trained VGG-16 network. Then it upsamples all features via transposed convolutional layers, and concatenates them for final prediction.

Comparing with the original implementation in [39], we set up a unified parallel structure for segmentation and registration, which enables more choices of transformed modalities other than the vessel maps. Besides, we replace the $\ell 1$ norm in the smoothness loss L_{sm} with $\ell 2$ norm to generate smoother registration fields, and remove the SSIM loss since it does not help in improving the registration performance.

2.4.3 Modality Transformer: Local Phase Signals

Instead of the vessel segmentation modality, we can also use the multi-scale local phase images, which is based on Monogenic signal [22], as a common modality to improve the registration performance in non-vessel areas. Previously, Li *et al.* [23] have shown the effectiveness of Monogenic local phase signals in a conventional multi-modal retinal registration pipeline.

In brief, the Monogenic signal is a multi-dimensional generalization of analytic signal. It can be computed by applying Riesz transformation on an input image [22], and the local phase term of the signal can be seen as the gradients of the image [70]. In order to extract image gradients in a certain range of scales (*i.e.*, frequencies), the input image is filtered with a 2D

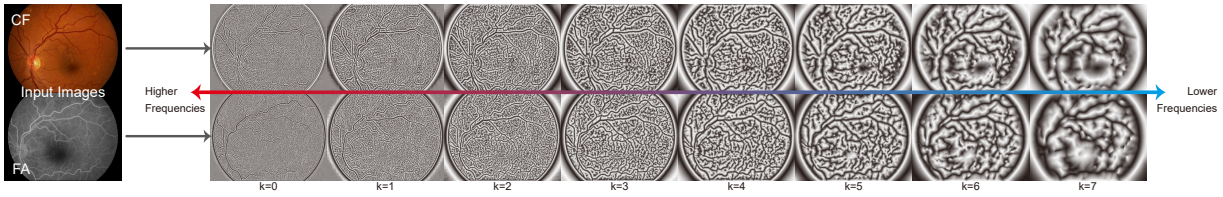


Figure 2.8: The phase signals of a multi-modal image pair are extracted with the Log-Gabor filters in Eq. (2.18) where $\sigma_0 = 0.55$ and $\omega_0 = 1/(5 \times 1.5^k)$, $k \in \{0, 1, ..., 7\}$.

log-Gabor band-pass filters prior to the Riesz transformation. The log-Gabor filter is given in frequency domain as

$$G(\omega) = \exp\left(- \frac{\left(\log(||\omega||/\omega_0)\right)^2}{2(\log(\sigma_0))^2} \right) \tag{2.18}$$

where ω_0 and σ_0 are center frequency and width of the filter which control the range of passed frequencies. After Riesz transformer, the 2-D local phase signal of the image I can be obtained as

$$\phi(I) = \arctan\left(\frac{\left(f_{o1}(I)^2 + f_{o2}(I)^2\right)^{1/2}}{f_e(I)} \right) \tag{2.19}$$

where $f_{o1}(I)$, $f_{o2}(I)$, and $f_e(I)$ are two odd parts and one even part of the Riesz transformation output from an filtered image.

In this chapter, we use K log-Gabor filters with $\sigma_0 = 0.55$ and $\omega_0 = 1/(5 \times 1.5^k), k = 0, 1, ..., K-1$ respectively to extract local phase maps at multiple scales. Fig. 2.8 shows an example of extracted local phase maps from multi-modal retinal images using the above log-Gabor filter settings. Filters with higher center frequencies ω_0 help extract finer details throughout the whole images, and results from lower frequency filters tends to focus on larger-scale patterns. Moreover, patterns in non-vessel areas are also extracted which can help increase their weights in the photometric consistency loss. Therefore, the alignment quality of the deformable registration network can be improved in the non-vessel areas.

2.5 Experiments

2.5.1 Settings

Dataset

We use two datasets, *i.e.*, CF-FA and JRC, for our experiments. CF-FA [71] is a public dataset captured in the modalities of CF and FA. It contains 59 pairs of retinal images of shape 720×576, 29 pairs of which are from healthy eyes and the rest show diseases. We take 30 pairs with odd index in the file names as the training set, and the remaining 29 for testing. The JRC dataset is collected by the Jacobs Retinal Center (JRC) at Shirley Eye Institute. It consists pairs of CF images of shape 3000×2672 and Infrared Reflectance (IR) images of shape 768×768 or 1536×1536. It has 530 pairs for training, 90 pairs for validation, and 253 pairs for testing. Especially, each image is graded by ophthalmologists as $\{$high / medium / low$\}$ according to its imaging quality, and as $\{$yes / no$\}$ based on the appearance of diseases.

Compared with the JRC dataset, the CF-FA dataset is less challenging since it shows crispy retinal patterns, and contains denser vessels and less diseases. An example of CF-FA images is shown in the first row of Fig. 2.2. However, the JRC dataset is more challenging, as its images often show sparser vessels, more diseases and unmatched structures. For example, in Fig. 2.2, the second example is considered as good quality, and the third example is graded as low quality due to the unwanted choroidal patterns and unfocused vessels in the CF image.

As an image preprocessing step, the images of the CF-FA dataset are expanded with zeros to 768×576, and the images of the JRC dataset are padded with zeros to the square shape and then resized to 768×768. To obtain the coarse alignment ground-truths for each image pair, we manually label three pairs of matching points and derive the ground-truth affine matrix $\mathbf{M}_{gt}$ based on the points' coordinates. In this chapter, we set CF as a source modality and FA/IR as a target, *i.e.*, CF images are warped towards a target modality.

Training and Testing Settings

For the outlier rejection network in the coarse alignment step, we set its input dimension as $N \leq 128$. We set $\lambda_c = 1$ and $\lambda_r = \lambda_D = 0.1$ in the loss function Eq. (2.10). Adam [72] optimizer is used for training with learning rate as $1e$-4. All the image coordinates are normalized into [-1, 1], and the ground-truth matrices $\mathbf{M}_{gt}$ are modified accordingly. The network is trained for 1000 epoches with batch size 32 on the JRC dataset. Due to small size of the CF-FA dataset, we take the model which is pre-trained on the JRC dataset and finetune it on the CF-FA dataset for 1000 epoches with batch size 30. The best checkpoint is selected based on the minimum Dice loss L_D on training set (CF-FA) or validation set (JRC).

For the fine alignment step, the networks are trained with Adam optimizer with learning rate as $1e$-3. We set $\lambda_{pc} = 1e$-3, $\lambda_{sm} = 5e$-4, $\lambda_{com} = 1e$-3, and $\lambda_{sty} = 1.0$ in Eq. (2.13) and (2.17). During training, two images of original size without any cropping are fed into the networks due to the requirement of style transfer loss, which takes up huge amount of GPU memory. Therefore, we set batch size to 1, and apply the same setting when training with local phase signals. The deformable networks are trained with 5000 (CF-FA) or 1500 (JRC) epoches, and the checkpoints with best $Dice_s$ value on the training set (CF-FA) or validation set (JRC) are selected for final evaluation. Two vessel segmentation images, *i.e.*, HRF-12h [67] in Fig. 2.7 (a) and DRIVE-28 [68] in Fig. 2.7 (c) from publicly available datasets, are selected as style targets for the CF-FA and JRC datasets respectively.

In addition, we employ data augmentation to train the fine alignment networks. First, training image pairs are set up based on $\mathbf{M}_{gt}$. For both datasets, coarsely aligned image pairs $\langle \text{STN}(I_{src}, \mathbf{M}_{gt}), I_{tgt} \rangle$ are used for training. For the CF-FA dataset, inversely aligned pairs $\langle I_{src}, \text{STN}(I_{tgt}, \mathbf{M}_{gt}^{-1}) \rangle$ are also included in the training set, which increases its size to 60. Next, the training pairs are augmented by random flipping (for both datasets) and rotation (for JRC only). Finally, random warping is applied on each image, *i.e.*, $2 \times 4 \times 3$ (for CF-FA) or $2 \times 4 \times 4$ (for JRC) arrays are first sampled from a normal distribution (mean 0, standard deviation 5 pixels),

then expanded to the image's resolution, and finally used to warp the image.

For evaluation on the JRC dataset, we separate the 253 test image pairs into 4 categories based on the gradings of imaging qualities and existence of diseases. In detail, the images are first categorized into two groups as High & Medium Quality and Low Quality. Then, the High & Medium Quality group is further divided into three sub-groups based on the number of images with diseases in each pair as in Table 2.2.

All networks are implemented in PyTorch and trained on GTX 1080 Ti GPU cards. During testing, all methods are tested on a desktop with a Intel i7-7700K CPU and a GTX 1080 Ti GPU card.

Evaluation Metrics

We adopt three evaluation metrics for registration quality assessment:

(a) $Dice \in [0, 1]$ is defined as the Dice coefficient of Eq. (2.7) which takes binary vessel segmentations from B-COSFIRE [73] as its inputs. The Dice coefficient is often used to evaluate retinal alignment quality (*e.g.*, [23]) when registration ground-truths are not available. It calculates the ratio of overlapping binary vessel areas to the sum of total vessel areas from both images. Larger Dice coefficients represent more overlapping area of vessels, which indicates better alignment quality. To extract binary vessels, we adopt B-COSFIRE [73] as the segmentation method. We keep the default settings in B-COSFIRE's codes except its segmentation threshold which is determined as follows. For the CF-FA dataset which shows better image qualities, two global thresholds are determined for each modality which maximizes the difference of *Dice* before and after warping based on the Phase + MIND [74] method in Section 2.5.3. For the JRC dataset, global thresholds lead to worse segmentation results (*i.e.*, too sparse or too dense vessels) due to the huge variance of image qualities, which impacts the alignment evaluation process. To ensure more reasonable segmentation results on the JRC test set, we estimate an individual threshold for each image which minimizes the style loss of Eq. (2.15) by comparing the thresholded result

with the style target DRIVE-28.

(b) $Dice_s \in [0,1]$ is defined as the Soft Dice of Eq. (2.8) which takes vessel probabilities from Frangi's [75] method as its inputs. Soft Dice is extended from the Dice coefficient and takes vessel probability maps. Similarly, it computes the ratio of vessel intersection to the sum of two vessel maps, and larger values indicate better alignment results. Frangi's [75] algorithm is used to extract vessel probabilities from retinal images. We first enhance a retinal image with CLAHE [76], then compute its vesselness map using Frangi's method, and finally rescale the vesselness map into $[0,1]$ based on its min & max values.

(c) #*success* is defined as the number of successfully aligned image pairs in each category. This metric is only used for the coarse alignment evaluation. The alignment success is achieved when

$$\max_{\mathbf{p}\in P} \left|\left| T\left(T(\mathbf{p}, \mathbf{M}_{gt}^{-1}), \mathbf{M}\right) \right|\right|_2 \leq \text{Threshold} \tag{2.20}$$

where $T(\cdot,\cdot)$ warps a coordinate $\mathbf{p}$ based on a transformation matrix, and P is a set of 6 correspondences for each image pair which is labeled by human. We empirically set Threshold $= 10$ *pixels*, *i.e.*, if all the source coordinates fall within the range of 10 pixels from their corresponding target coordinates after transformation, it is considered a success.

Fig. 2.9 shows an example of *Dice* and $Dice_s$ calculation. The overlapping maps of their extracted vessel binaries or probabilities are plotted before and after registration. In the left column, some tiny vessels and non-vessel structures from the source image (red) are missed in *Dice* due to the binary thresholding. But those structures are preserved as probabilities in $Dice_s$ (right column), and can be included in the registration quality evaluation. Therefore, *Dice* mainly assesses the alignment quality of prominent vessels, while $Dice_s$ pays attention to both major and tiny retinal structures.

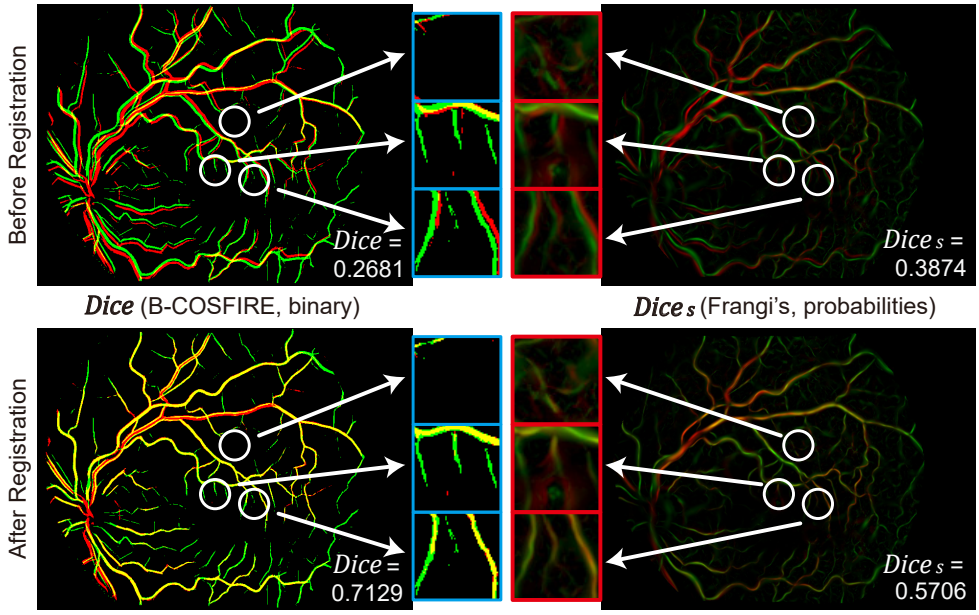


Figure 2.9: A comparison of *Dice* and $Dice_s$ over a same pair of input images before and after registration. Red and green areas indicate extracted vessel binaries or probabilities from source and target image respectively, and yellow areas indicate their overlapping parts. White circles point to tiny vessels which are missed in *Dice* but maintained in $Dice_s$. Pixel intensities in red boxes are enhanced for visual inspection.

2.5.2 Results on Two-Step Coarse-to-Fine Registration

For comparison, four groups of methods/results are set up for coarse-to-fine registration evaluation for the quantitative and qualitative comparison in Table 2.2 and Fig. 2.10, respectively. The groups are:

(a): Input images without any warping.

(b): An improved version of a conventional two-step registration pipeline [23], where the time-consuming matching algorithm is replaced by a much faster one [42]. Specifically, local phase signal is adopted to help its feature-based coarse alignment step (denoted as Phase [22] - HoG [77] - RANSAC [43]), and then MIND[1] [74] is used for fine alignment.

(c): A state-of-the-art optical flow network IRR-PWC [56] with supervised training. Since

[1]Only its deformable registration part is used in this chapter.

coarse-to-fine registration pipelines based on fully CNN are unavailable, a pre-trained IRR-PWC is adopted and finetuned on multi-modal retinal images. The network takes an image pair as input and estimates a registration field map which warps the source image in one step. Specifically, two types of inputs are fed into IRR-PWC, *i.e.*, original retinal images, or predicted vessel maps from our segmentation networks which mimics the optical flow estimation process. For each type of input, two models are trained on different optical flow ground-truths, and the better one is used for evaluation. The ground-truth optical flows are generated based on either $\mathbf{M}_{gt}$ only, or a combination of $\mathbf{M}_{gt}$ and our best deformable alignment result (*i.e.*, results in Section 2.5.3).

(d): Our proposed pipeline, where we denote the coarse alignment network as CoarseNet, the fine alignment network with the modality transformer for the vessel segmentation as Seg-DeformNet, and the fine alignment network with the modality transformer for the local phase signals as Phase-DeformNet.

Table 2.2 shows the quantitative evaluation results for the JRC (blue columns) and CF-FA (red column) datasets. On the JRC dataset, the proposed two-step methods in group (d) achieve the best performance in both *Dice* and $Dice_s$ metrics. Moreover, the conventional two-step method (b2) outperforms the single-step networks with supervised training in group (c), which shows that two-step approaches are more effective for multi-modal registration. When we compare *Dice*/$Dice_s$ of the proposed two-step networks in group (d) with the conventional method (b2) for each category from left to right, the advantage of our method gradually increases as the images become more challenging. For example, the gains of row (d3) over (b2) from left to right are $0.0941/0.0529$ for No Disease, $0.1093/0.0618$ for 1 Disease, $0.1249/0.0658$ for 2 Diseases, and $0.1724/0.0903$ for Low Quality images. This result demonstrates that our two-step methods are more robust to multi-modal images with more disease lesions and lower image qualities.

The last column in Table 2.2 shows the results for the CF-FA dataset. The rankings of *Dice*/$Dice_s$ values of groups remain similar with that on the JRC dataset. The advantage of our networks (d) over the conventional method (b2) (*i.e.*, difference between row (d3) and (b2)) is

Table 2.2: Average *Dice*/*Dice*$_s$ Values for Two-Step Registration on the JRC and CF-FA Datasets

Method				JRC Dataset (253 images)					CF-FA Dataset (29 images)
Group	Row #	Coarse Alignment	Fine Alignment	High & Medium Quality (203)			Low Quality (50)	Overall (253)	
				No Disease (111)	1 Disease (42)	2 Diseases (50)			
(a)	(a1)	Before registration		0.0734 / 0.2536	0.0754 / 0.2565	0.0732 / 0.2666	0.0749 / 0.2679	0.0740 / 0.2595	0.1012 / 0.2582
(b)	(b1)	[23] (Phase-HoG-RANSAC)	N/A	0.4160 / 0.4461	0.4181 / 0.4490	0.3431 / 0.4175	0.1792 / 0.3170	0.3551 / 0.4154	0.5504 / 0.5050
	(b2)		MIND [74]	0.4894 / 0.5047	0.4726 / 0.4961	0.4094 / 0.4692	0.2166 / 0.3435	0.4169 / 0.4644	0.6289 / 0.5291
(c)	(c1)	IRR-PWC [56] (input: vessel)		0.3754 / 0.4007	0.3479 / 0.3916	0.3135 / 0.3866	0.2332 / 0.3384	0.3305 / 0.3841	0.2959 / 0.3910
	(c2)	IRR-PWC [56] (input: image)		0.3505 / 0.3913	0.3369 / 0.3880	0.3178 / 0.3894	0.2590 / 0.3523	0.3237 / 0.3826	0.2251 / 0.3535
(d) **Ours**	(d1)	**CoarseNet**	N/A	0.5701 / 0.5129	0.5623 / 0.5052	0.4949 / 0.4793	0.3331 / 0.3864	0.5071 / 0.4800	0.5902 / 0.5204
	(d2)		**Seg-DeformNet**	**0.6040** / 0.5431	**0.6034** / 0.5384	**0.5549** / 0.5185	**0.4139** / **0.4356**	**0.5566** / 0.5162	**0.6812** / 0.5394
	(d3)		**Phase-DeformNet**	0.5835 / **0.5576**	0.5819 / **0.5534**	0.5343 / **0.5350**	0.3890 / 0.4338	0.5350 / **0.5280**	0.6678 / **0.5412**

Table 2.3: Average *Dice*/*Dice*$_s$ Values for Fine Alignment on the JRC and CF-FA Datasets

Fine Alignment Method	JRC Dataset					CF-FA Dataset
	High & Medium Quality			Low Quality	Overall	
	No Disease	1 Disease	2 Diseases			
$\mathbf{M}_{gt}$ + Random Warping	0.3459 / 0.3920	0.3320 / 0.3904	0.3048 / 0.3851	0.2412 / 0.3430	0.3148 / 0.3807	0.2744 / 0.3934
MIND [74]	-	-	-	-	-	0.6019 / 0.5269
Phase [22] + MIND [74]	0.5426 / 0.5196	0.5302 / 0.5124	0.4794 / 0.4889	0.3527 / 0.4006	0.4905 / 0.4888	0.6178 / 0.5303
VoxelMorph (NCC) [60]	-	-	-	-	-	0.6408 / 0.5358
Zhang *et al.* [39]	-	-	-	-	-	0.6546 / 0.5398
Seg-DeformNet (Ours)	**0.5970** / 0.5355	**0.5910** / 0.5320	**0.5426** / 0.5120	**0.4173** / **0.4237**	**0.5497** / 0.5082	**0.6692** / **0.5442**
Phase-DeformNet (Ours)	0.5675 / **0.5442**	0.5669 / **0.5427**	0.5065 / **0.5197**	0.3888 / 0.4225	0.5201 / **0.5151**	0.6487 / 0.5435

*VoxelMorph failed in multiple trials on the JRC dataset, by predicting either NaN values or extremely large displacement fields.

reduced to $0.0389/0.0121$ since the CF-FA dataset contains fewer challenge cases than the JRC dataset. Nevertheless, the single-step optical flow networks with supervised training in group (c) have the lowest performance.

In Fig. 2.10, we compare qualitative results on two image pairs from JRC dataset where a normal pair and a diseases pair are shown in example 1 and 2 respectively. In both examples, our methods can correctly and accurately align most vessels and disease patterns. In contrast, the other methods fail in at least one of the examples, *e.g.*, Phase - HoG - RANSAC + MIND of (b2) fails in the example 2 to align vague vessels (box 2) and lesions (box 3) and the IRR-PWC's results contain obvious misalignment errors in all examples, as indicated by red arrows.

Apart from the above comparison, within each group (b) and (d) in Table 2.2, the additional fine alignment steps of (b2)/(d2)/(d3) are able to improve the registration performance over coarse alignment results of (b1)/(d1). This is also demonstrated in red circles of Fig. 2.10, where misalignments from CoarseNet of (d1) are corrected by fine alignment networks in (d2)/(d3). This result presents clearly that the refinement step can help to correct errors from the first step and thus increase the robustness of the registration pipeline. Therefore, two-step coarse-to-fine structures achieve better results comparing to single-step ones for multi-modal retinal registration.

2.5.3 Ablation Study on Deformable Registration Networks

In this section, we investigate the performance of the proposed fine alignment networks on both datasets by comparing it with other methods. Moreover, we further analyze several factors that influence its registration quality. Testing image pairs aligned by $\mathbf{M}_{gt}$ are adopted for evaluation. All the testing images are preprocessed by the identical augmentation procedure in fine alignment network training which is described in Section 2.5.1. Therefore, the size of the CF-FA test set is doubled to 58 in this section. The random warping flows applied on the input images are fixed for all methods.

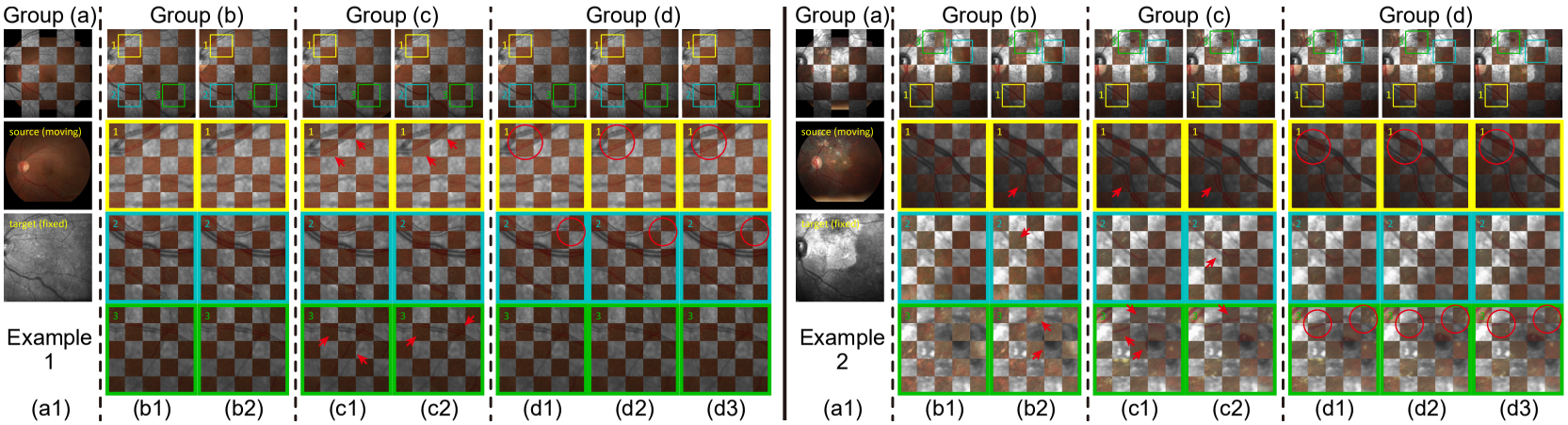
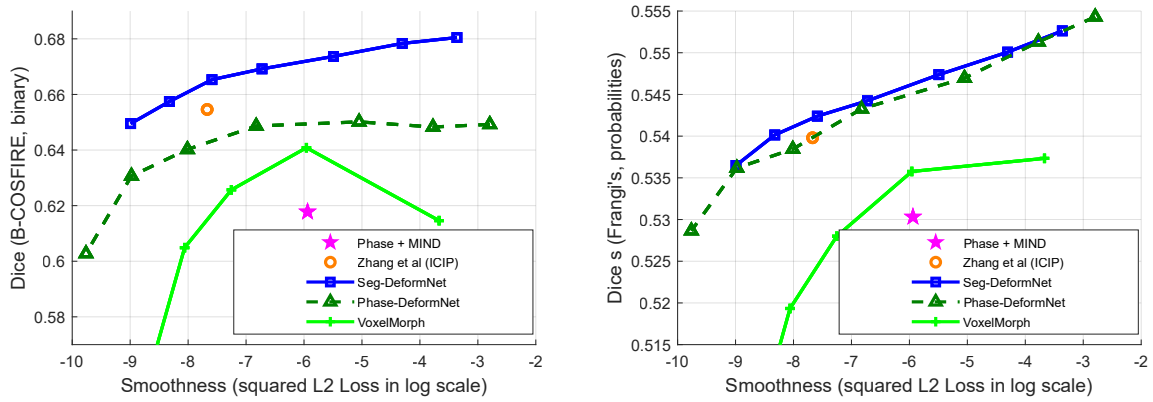


Figure 2.10: Two-step registration results on two examples from the JRC dataset, where source and target images are displayed interlaced as small grids. From top to bottom lines are: (a1) Input images, (b1) Phase-HoG-RANSAC (coarse only), (b2) Phase-HoG-RANSAC + MIND (two-step), (c1) IRR-PWC (input: vessel) (c2) IRR-PWC (input: image), (d1) CoarseNet (coarse only), (d2) CoarseNet + Seg-DeformNet (two-step), (d3) CoarseNet + Phase-DeformNet (two-step). In each example, the leftmost column shows complete images, and the following three columns show magnified details in the yellow, blue and green boxes denoted by numbers. Red Arrows point to misalignment of retinal structures. Red circles show improved alignment results by the second step in a two-step registration pipeline over its global alignment results.

Fine Alignment Evaluation

Table 2.3 shows the deformable registration results for the JRC and CF-FA datasets. The proposed deformable networks outperform the conventional method Phase + MIND in all cases, showing the advantages of CNN in this task. In addition, our proposed networks achieve better performance than another unsupervised network, VoxelMorph [60], which uses local NCC as the similarity metric in training. This shows the advantages of transformed modalities (vessel maps and phase signals) over conventional image-based similarity metrics in the deformable registration task on multi-modal retinal images.

When comparing Phase-DeformNet and Seg-DeformNet on the JRC dataset in the blue columns, Phase-DeformNet ranks highest in $Dice_s$ in most categories except the low quality images, while Seg-DeformNet performs best in *Dice* across all groups. However, on the CF-FA dataset in the red column, Seg-DeformNet achieves the best performance in both $Dice_s$ and *Dice*, although its advantage in $Dice_s$ is marginal (*i.e.*, +0.0007). Similar relations of two methods are also observed in Table 2.2, where the gap of $Dice_s$ values between Seg-DeformNet and Phase-DeformNet is smaller on the CF-FA dataset (*i.e.*, -0.0018) than that on the JRC dataset (0.0118 in the Overall column). It might result from the different supervision signals in the content loss (Eq. (2.11)) and different characteristics of the datasets. Specifically, Seg-DeformNet is trained by optimizing the alignment of extracted vessels, and cannot directly align non-vessel areas if their segmentation predictions are zero. Thus, it tends to get higher values in *Dice* (*i.e.*, the overlapping degree of prominent vessels) but lower values in $Dice_s$ (which puts more emphasis on non-vessel areas). Since Phase-DeformNet is trained by aligning local phase patterns which distribute over the whole images, it tends to achieve better performance in non-vessel areas, *i.e.*, higher $Dice_s$. On the other hand, images in the CF-FA dataset have denser vessels and less non-vessel patterns (*e.g.*, diseases) than those in the JRC dataset. Therefore, it is possible for Seg-DeformNet to have minor margin over Phase-DeformNet on the CF-FA dataset by only aligning vessel patterns.



(a) *Dice* to L_{sm}.

(b) $Dice_s$ to L_{sm}.

Figure 2.11: Deformable registration performance on the CF-FA dataset for networks trained with various λ_{sm}. Each data point indicates *Dice*/$Dice_s$ and L_{sm} values measured on the test set. The number besides each point indicates the corresponding λ_{sm}.

Smoothness Weight

We analyze the influence of different smoothness factor λ_{sm} in Eq. (2.13) and (2.17) on the registration results. Theoretically, higher λ_{sm} makes it more difficult to correct abrupt misalignments in small areas and reduces alignment quality because of competitions between photometric consistency loss L_{pc} (Eq. (2.11)) and smoothness loss L_{sm} (Eq. 2.12). In this experiment, we set $\lambda_{sm} \in \{5e\text{-}3, 2e\text{-}3, 1e\text{-}3, 5e\text{-}4, 2e\text{-}4, 1e\text{-}4, 5e\text{-}5\}$ for both Seg-DeformNet and Phase-DeformNet and train them on the CF-FA dataset where other settings remain unchanged. The evaluation results on the test set are displayed in Fig. 2.11 as the relations between *Dice*/$Dice_s$ values and L_{sm} (Eq. (2.12)). As can be seen from Fig. 2.11 (b), $Dice_s$ of two methods keep increasing as λ_{sm} decreases, which complies with the theoretical analysis. Moreover, from Fig. 2.11 (a), the trend of increase in *Dice* for Phase-DeformNet stops below $\lambda_{sm} = 5e\text{-}4$, which implies that the network cannot make improvement in aligning vessels by reducing λ_{sm}. Therefore, we choose $\lambda_{sm} = 5e\text{-}4$ as our setting in Section 2.5.2.

Fig. 2.11 also plots the results of the conventional method (*i.e.*, Phase + MIND) and our previous network [39]. The proposed Seg-DeformNet (the blue point at $\lambda_{sm} = 1e\text{-}3$) has better

alignment quality than our previous network [39] (the orange point) at a similar smoothness level. Besides, both the proposed Seg-DeformNet and Phase-DeformNet achieve higher $Dice/Dice_s$ values than the conventional method (the purple star), which shows the strengths of DNN on this task.

Diffeomorphic Property of Deformation Fields

Diffeomorphic registration is another research topic in medical image registration and has been incorporated in recently proposed networks [78–80]. It aims to obtain a topology-preserving and invertible transformation that enforces one-to-one mapping. Nevertheless, it is not a major concern or contribution in this work. Therefore, we only measure the diffeomorphic property of our predicted registration fields during testing.

The diffeomorphic property of a registration field can be analyzed by the determinant of Jacobian matrix, which is defined on each pixel (x,y) as

$$|J_F(x,y)| = \begin{vmatrix} \frac{\partial_{F_{0,\cdot,\cdot}}(x,y)}{\partial x} & \frac{\partial_{F_{0,\cdot,\cdot}}(x,y)}{\partial y} \\ \frac{\partial_{F_{1,\cdot,\cdot}}(x,y)}{\partial x} & \frac{\partial_{F_{1,\cdot,\cdot}}(x,y)}{\partial y} \end{vmatrix}, \tag{2.21}$$

where $F_{0,\cdot,\cdot}$ and $F_{1,\cdot,\cdot}$ are the maps of displacement vectors in two directions. If $|J_F(x,y)| < 0$, the deformation at (x,y) fails to preserve the same warping orientation as its neighbors, which is unfavored. In this work, we compute $|J_F(x,y)|$ for all predicted registration fields on the test sets, and count the number of pixels with negative determinant values. On the CF-FA dataset, the percentage of pixels with negative determinant values are 0.0002% for Seg-DeformNet and 0% for Phase-DeformNet. On the JRC dataset, the values become 0.0044% for Seg-DeformNet and 0% for Phase-DeformNet. These numbers show that only a small portion of pixels in the predicted registration fields have negative values. Therefore, the diffeomorphic property is mostly preserved by our proposed deformable networks.

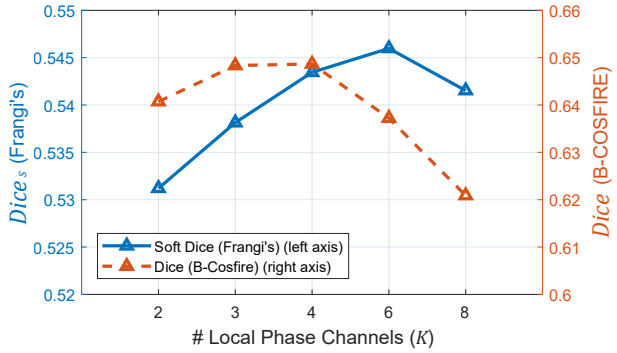


Figure 2.12: Registration performance of Phase-DeformNet on the CF-FA dataset when trained with different number of local phase channels K.

Number of Local Phase Image Channels

The relation of Phase-DeformNet's performance with regard to the number of channels K in local phase signal is shown in Fig. 2.12. In theory, networks trained with small K can only see detailed information (high-frequency patterns) in alignment. Increasing K includes more low-frequency information for registration. As shown in Fig. 2.12, *Dice* peaks on both $K = 3$ and $K = 4$ and start to decrease sharply from $K = 6$. $Dice_s$ keeps increasing from $K = 2$ to $K = 6$ and decrease afterwards. In order to achieve a balance between *Dice* and $Dice_s$, we select $K = 4$ for Phase-DeformNet in our experiments.

Style Target Choices

The influences over Seg-DeformNet's performance from different style targets are demonstrated in Table 2.4. In details, three different segmentation maps are selected from HRF [67] and DRIVE [68] datasets, as shown in Fig. 2.7 where the vessel density of HRF-12h, DRIVE-23, and DRIVE-28 is dense, sparse, and in-between, respectively. On the CF-FA dataset, the network trained with HRF-12h achieves the best performance. On the JCR dataset, DRIVE-28 helps the network to obtain the highest *Dice*/$Dice_s$ values. We attribute this performance's variations of a certain style target to the similarities between the style target and the datasets' images. Since images in the CF-FA dataset generally have good quality and dense vessel structures, a style

Table 2.4: Average *Dice*/*Dice*$_s$ Values of Seg-DeformNet on the JRC and CF-FA Datasets Trained with Different Style Targets

Style Target Image	JRC	CF-FA
HRF-12h [67]	0.5088 / 0.5056	**0.6692 / 0.5442**
DRIVE-23 [68]	0.5290 / 0.4963	0.6345 / 0.5353
DRIVE-28 [68]	**0.5497 / 0.5082**	0.6559 / 0.5416

target with denser vessels (*e.g.*, HRF-12h) might achieve better vessel segmentation, and thus leads to better alignment results. However, the images in the JRC dataset has relatively sparse vessel densities and also show lower image quality. Therefore, the best result is achieved with the sparser vessel style image DRIVE-28 instead of HRF-12h.

2.5.4 Ablation Study on Coarse Alignment

In this section, we investigate the performance of the coarse alignment step on the JRC dataset. We set up additional groups of methods and training settings in Table 2.5 as follows:

(e): DRMIME [81] and the AffineNet in DLIR [14] adopt image-based similarity metrics. DRMIME is an iterative optimization method for multi-modal registration based on the MI metric. Specifically, it computes approximate MI values via MIME (mutual information neural estimation) [83], and sets up input image pyramids for registration. In our experiments, we use five pyramid layers and 20% random sampling for registration, and keep the other settings unchanged. In order to remove most contours in CF and reduce initial scale differences between both modalities, we use a different image preprocessing step, *i.e.*, a CF images is downsampled by $1/3$ from its original resolution, and then its center 768×768 area is cropped for optimization.

DLIR is a coarse-to-fine cascade pipeline that connects an affine registration network and multiple deformable networks. The deformable networks are trained based on the outputs of previous networks. Negative NCC is used as the loss function for unsupervised training. We only implement the affine network instead of the full pipeline.

(f) and (g): We investigate the influences on the registration performance from different com-

Table 2.5: Coarse Alignment Performance on the JRC Dataset: Average *Dice*/*Dice*$_s$ (#*Success*)

Group	Row #	Coarse Alignment Method	High & Medium Quality (203)			Low Quality (50)	Overall (253)
			No Disease (111)	1 Disease (42)	2 Diseases (50)		
(a)	(a1)	Before registration	0.0734 / 0.2536	0.0754 / 0.2565	0.0732 / 0.2666	0.0749 / 0.2679	0.0740 / 0.2595
(b)	(b1)	[23] (Phase-HoG-RANSAC)	0.4160 / 0.4461 (55)	0.4181 / 0.4490 (23)	0.3431 / 0.4175 (19)	0.1792 / 0.3170 (5)	0.3551 / 0.4154 (102)
(e)	(e1)	DRMIME [81]	0.1003 / 0.2720 (1)	0.0906 / 0.2706 (0)	0.0817 / 0.2753 (0)	0.0832 / 0.2706 (0)	0.0916 / 0.2721 (1)
	(e2)	AffineNet of DLIR [14]	0.0805 / 0.2558 (0)	0.0749 / 0.2591 (0)	0.0799 / 0.2750 (0)	0.0811 / 0.2701 (0)	0.0795 / 0.2630 (0)
(f) Input+ SIFT+ RANSAC	(f1)	IR (MUNIT [82])	0.2217 / 0.3227 (21)	0.1850 / 0.2914 (4)	0.1298 / 0.2591 (3)	0.1133 / 0.2520 (0)	0.1760 / 0.2910 (28)
	(f2)	B-Cosfire [73]	0.2346 / 0.3269 (30)	0.2525 / 0.3379 (13)	0.1827 / 0.3008 (9)	0.0962 / 0.2494 (2)	0.2000 / 0.3083 (54)
	(f3)	Phase [22]	0.1473 / 0.2498 (11)	0.1345 / 0.2705 (2)	0.1066 / 0.2558 (2)	0.0746 / 0.2275 (0)	0.1228 / 0.2500 (15)
	(f4)	SegNet (Ours)	0.1968 / 0.3090 (21)	0.2005 / 0.3073 (10)	0.1344 / 0.2849 (6)	0.0804 / 0.2372 (1)	0.1621 / 0.2898 (38)
(g) Input+ SuperPoint+ RANSAC	(g1)	IR (MUNIT [82])	0.4489 / 0.4553 (79)	0.4363 / 0.4426 (28)	0.3517 / 0.4089 (25)	0.1552 / 0.2955 (8)	0.3696 / 0.4124 (140)
	(g2)	B-Cosfire [73]	0.4310 / 0.4358 (70)	0.4000 / 0.4178 (24)	0.3324 / 0.3901 (23)	0.2259 / 0.3147 (11)	0.3659 / 0.3999 (128)
	(g3)	Phase [22]	0.5006 / 0.4808 (89)	0.4447 / 0.4532 (30)	0.3942 / 0.4194 (29)	0.1371 / 0.2853 (4)	0.3984 / 0.4254 (152)
	(g4)	SegNet (Ours)	0.5189 / 0.4859 (92)	0.5072 / 0.4769 (34)	0.3985 / 0.4162 (31)	0.2254 / 0.3062 (13)	0.4352 / 0.4351 (170)
(h)	(h1)	CoarseNet (w/o L_D)	0.5693 / 0.5141 (106)	0.5623 / **0.5085** (40)	0.4942 / 0.4806 (43)	**0.3370 / 0.3866** (21)	**0.5074** / 0.4813 (210)
	(h2)	CoarseNet (w/o L_c)	0.5582 / 0.5058 (104)	0.5402 / 0.4934 (38)	0.4652 / 0.4649 (34)	0.2862 / 0.3653 (16)	0.4831 / 0.4679 (192)
	(h3)	CoarseNet (w/o L_r)	0.5698 / **0.5149** (107)	**0.5630** / 0.5077 (41)	**0.4976 / 0.4826** (41)	0.3260 / 0.3848 (23)	0.5062 / **0.4816** (212)
(d)	(d1)	CoarseNet	**0.5701** / 0.5129 (107)	0.5623 / 0.5052 (41)	0.4949 / 0.4793 (46)	0.3331 / 0.3864 (24)	0.5071 / 0.4800 (218)

Table 2.6: Coarse Alignment Performance of CoarseNet on the JRC Dataset, Trained with Corrupted Ground-Truths: Average *Dice*/*Dice*$_s$ (#*Success*)

Corruption Level of Ground-Truth	Training Settings	High & Medium Quality (203)			Low Quality (50)	Overall (253)
		No Disease (111)	1 Disease (42)	2 Diseases (50)		
	Before registration	0.0734 / 0.2536	0.0754 / 0.2565	0.0732 / 0.2666	0.0749 / 0.2679	0.0740 / 0.2595
5% Corruption	w/o L_D	0.5218 / 0.4856 (98)	0.5190 / 0.4826 (35)	0.4374 / 0.4544 (32)	0.2762 / 0.3550 (15)	0.4561 / 0.4531 (180)
	w/ L_D, $\lambda_D = 0.1$	0.5345 / 0.4929 (102)	0.5268 / 0.4871 (37)	0.4353 / 0.4532 (34)	0.2864 / 0.3632 (15)	0.4646 / 0.4584 (188)
10% Corruption	w/o L_D	0.4529 / 0.4471 (69)	0.4424 / 0.4401 (26)	0.3772 / 0.4207 (23)	0.2281 / 0.3373 (10)	0.3918 / 0.4190 (128)
	w/ L_D, $\lambda_D = 0.1$	0.4690 / 0.4547 (78)	0.4308 / 0.4346 (24)	0.3730 / 0.4149 (25)	0.2290 / 0.3354 (7)	0.3963 / 0.4208 (134)
	w/ L_D, $\lambda_D = 1$	0.4999 / 0.4742 (83)	0.4818 / 0.4634 (34)	0.4127 / 0.4393 (33)	0.2456 / 0.3453 (11)	0.4294 / 0.4400 (161)

binations of input image modalities and features. We set up four types of mono-modal inputs, i.e. IR images, B-Cosfire [73] vessel segmentation maps, averaged local phase maps ($K = 4$), and the vessel maps from our segmentation network (SegNet). Especially, we trained a MUNIT [82] model on the training set to translate all CF images into the IR modality. We do not use the opposite translation path (*i.e.*, IR to CF) because of lower quality in the synthesized CF images. Besides, we use two feature detectors and descriptors, *i.e.*, SIFT [42] and the pre-trained SuperPoint [20] network.

(h): We retrain the outlier rejection network in three additional settings, with each ignoring one loss term in Eq. (2.10).

Groups **(a), (b) and (d)** are from Table 2.2.

Coarse Alignment Evaluation

In Table 2.5, we compare our proposed network (d1) with three other methods, *i.e.*, (b1) a feature based conventional registration pipeline [23], (e1) an iterative optimization method [81] based on MI, and (e2) an unsupervised affine network [14] based on NCC. As observed, our proposed network has a large advantages in $Dice/Dice_s$ values over the compared methods. Moreover, the methods (e1) and (e2) which use image-based similarity metrics fail in most cases, *i.e.*, only 1 and 0 successful alignment respectively, which has been implied in Fig. 2.2. This indicates that multi-modal retinal image registration is a very challenging task for intensity-based methods.

Choice of Input Modalities and Features

When combined with SIFT features (Group (f) in Table 2.5), the overall registration performance remains at lower levels, with B-Cosfire (f2) achieving the best performance. However, the SuperPoint network (Group (g)) boosts the registration performance over SIFT by a large

margin, especially for the modalities of Phase maps (g3) and the vessel maps by our segmentation networks (g4). Finally, the overall optimal performance is achieved by (g4) SegNet + SuperPoint, which is also adopted in our coarse alignment step.

Loss Terms for Outlier Rejection Network

In Table 2.5 Group (h), we investigate the influences of different loss terms in Eq. (2.10) on the outlier rejection performance. By comparing the various settings in (h) with (d1) which is trained with the complete loss function, it shows that the largest performance drop is triggered by removing the classification loss, while ignoring the other two terms has less impact on the alignment quality. This demonstrates the major contribution from the classification loss in training the outlier rejection network for our task.

In Table 2.6, we further investigate the value of the proposed Dice loss L_D in the cases of training with uncleaned ground-truths. We adjust each element in the ground-truth matrices $\mathbf{M}_{gt}$ by random percentages sampled in $[-5\%, 5\%]$ or $[-10\%, 10\%]$ to simulate polluted labels. Then we train the outlier rejection network on these labels under various settings. As observed, when trained with uncleaned labels, the settings using L_D show better alignment results than the ones without L_D. Besides, increasing the weight λ_D for Dice loss can further improve the performance.

2.5.5 Runtime Analysis

A disadvantage of our proposed network is that, the segmentation networks take more time and large GPU memory in training. This is mainly the result of the style loss computation, which compares style features between two complete segmentation maps, and thus requires the inputs of the complete retinal images. When training Seg-DeformNet on the JRC dataset, the network takes 7.9 GB in GPU memory and 9 days for training. In the other hand, the Phase-DeformNet takes 1.8 GB of GPU memory and 4 days for training, since it does not perform segmentation. In addition, the outlier rejection network takes 4.5 GB GPU memory and 7 hours for training.

Table 2.7: Testing Runtime of Each Method on the JRC Dataset

Method	Registration Step	Platform	Time/ Pair	GPU Memory
DRMIME (500 iters)	Coarse	PyTorch	49.5s	1.2 GB
Phase-HoG-RANSAC	Coarse	Matlab	1.51s	
Phase+MIND	Fine	Matlab	41.9s	
IRR-PWC	Two-Step	PyTorch	0.264s	1.3 GB
Ours	Two-Step	PyTorch	0.705s	7.8 GB
–Segmentation			0.525s	
–Feature Detection & Description			0.141s	
–Outlier Rejection			0.0277s	
–Deformable Registration			0.0117s	

There are 0.26M trainable parameters in each segmentation network (in addition to the pre-trained VGG-16 layers with 14.7M parameters). Besides, the networks for feature detection and description, outlier rejection, and deformable registration have 1.30M, 1.58M and 1.53M parameters respectively. Table 2.7 shows the per-pair testing time of our network on the 768×768 images pairs from the JRC dataset. Our proposed two-step pipeline takes less than one second for prediction, which is much faster than the conventional methods.

2.6 Conclusion

In this chapter, we set up a two-step coarse-to-fine CNN-based registration algorithm for multi-modal retinal images. In the coarse registration step, an accurate transformation matrix is estimated for an image pair through extracting vessels, finding features and eliminating outlier matchings with three consecutive networks. The fine alignment step further improves alignment quality by estimating a pixel-wise registration map using a deformable registration network. To train the deformable networks, we propose to transform multi-modal images into a common modality to fulfill the color consistency requirements in the unsupervised training scheme. We propose to train vessel segmentation networks via style loss, which can benefit both coarse and fine alignment steps. Experiment results show that our method achieve the state-of-the-art registration result in both quantitative and quality measurements.

In the future, we would like exploit the potential of our proposed method on aligning more challenging retinal images *e.g.*, multi-phase images, images with large scale variations, ultra wide field images, etc. In addition, Generative Adversarial Networks (GAN) that disentangle image content and styles, *e.g.*, [82], have shown feasibilities in aligning [84] multi-modal medical images. By incorporating GAN as modality transformers into our unsupervised joint learning scheme may further unveil the potentials of both methods.

Chapter 2, in full, is a reprint of the material as it appears in IEEE Transactions on Image Processing, 2021. Junkang Zhang; Yiqian Wang; Ji Dai; Melina Cavichini-Cordeiro; Dirk-Uwe G. Bartsch; William R. Freeman; Truong Q. Nguyen; Cheolhong An, IEEE, 2021. The dissertation author was the primary investigator and author of this material.

3 Multimodal Global Registration between Ultra-Widefield and Narrow Angle Retinal Images via Distortion Correction Network

3.1 Introduction

Retinal image alignment plays an important role in the screening and diagnosis of retinal diseases as well as management of patient information. Conventional Narrow Angle (NA) fundus cameras can capture high-resolution details in a small retinal area. However, it is demanding for patients and doctors to image larger retinal fields, because it takes much more time to capture multiple images to cover peripheral retina and requires more skills for doctors to manipulate the fundus instrument. Recently, Ultra Wide Field (UWF) cameras have also become popular choices for retinal imaging, because they can capture a much larger view of retina in a single fast shot, which makes them more convenient alternatives for disease screening. With the alignment of UWF and NA images, doctors can acquire a larger view of the retina as well as detailed information in disease areas, which will ease burdens for both doctors and patients. Furthermore, by overlaying multi-modal retinal images captured under lights of multiple wavelengths, doctors

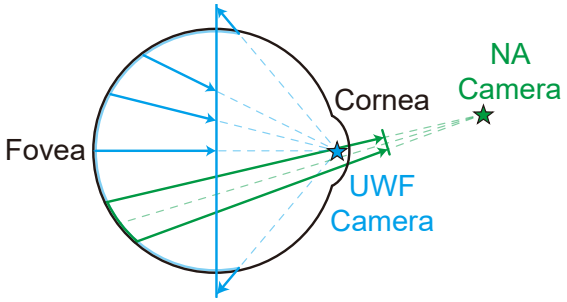


Figure 3.1: Simplified illustrations of UWF and NA retinal cameras. Blue and green arrows show the projection paths of 3D spherical points onto 2D UWF and NA imaging planes respectively.

can also obtain more comprehensive information and detailed analysis for the same lesions, as the retinal structures have different appearance in multiple modalities.

There has been extensive research on multi-modal retinal image registration [24–35, 38, 39, 85, 86]. However, the alignment of UWF and NA retinal images remains unexploited. Most proposed methods were designed for NA-to-NA image alignment via 2D-to-2D transformations, which have yet to be tested on UWF image alignment. Meanwhile, there remains an unsolved challenge in the UWF-to-NA alignment process, *i.e.*, misalignment due to nonlinear distortions in the UWF images. In detail, in order to achieve a wider field of view through the pupil, the UWF camera sets its view point closer to the eye ball than most NA cameras, as shown in Fig. 3.1. This design leads to larger perspective distortions in peripheral areas due to the 3D spherical shape of eyeball, where retinal structures closer to the camera appear larger in scale than those distant ones. Therefore, the UWF-to-NA alignment performance will be degraded in two aspects. On one hand, it is challenging to model these distortions by a global 2D-to-2D transformation process, which will have many misalignments when overlaying NA images onto the UWF peripheral areas. On the other hand, in feature-based alignment methods, it will be more difficult to correctly match the extracted features from the heavily-distorted UWF and less-distorted NA images, which will further reduce the alignment accuracies.

To solve this challenge, the eyeball's 3D shape prior must be considered in the retinal

Figure 3.2: Comparison of UWF-to-NA and NA-to-NA alignment via perspective transformation. Left side shows the alignment of an Optos UWF ColorMap (floating) and a Spectralis MultiColor (NA, fixed) images. Right side shows the alignment of a Color Fundus (NA, floating) and a Spectralis Infrared (NA, fixed) images. The results are shown as the overlapping vessel maps of both images. Red and green vessels are from floating and fixed images respectively, and yellow areas indicate their overlapping parts.

image alignment process. Multiple eyeball shape assumptions have been incorporated into NA-to-NA retinal image alignment, including spherical [87,88], quadratic [88], and ellipsoidal priors [89], etc. In these cases, the alignment task can be handled as a pose estimation problem. Nevertheless, existing methods only align NA images which is a easier task, because the distortions between neighboring images are much less critical due to their smaller retinal coverage and thus the images have more similar camera poses *e.g.*, the NA-to-NA alignment of Color Fundus and Infrared retinal images shown on the right side of Fig. 3.2. On the other hand, between NA and UWF images, the distortions could be a critical issue in alignment if the NA's camera pose departs sharply from the camera pose of UWF, *e.g.*, on the left side of Fig. 3.2, the UWF-to-NA alignment quality is heavily degraded by various levels of image distortions.

In this paper, we propose a distortion correction algorithm on UWF images which incorporates the 3D spherical prior for the UWF-to-NA multi-modal image alignment. First, we assume that the eyeball is a pure sphere, and the optical axis of the UWF camera goes from the pupil center to the fovea center, *i.e.*, an ideal 3D eyeball shape and a fixed pose for the UWF camera. Then, we set up reprojection functions to remap the UWF image based on a similar camera pose of the NA image to be aligned. The NA camera pose is described by 5 parameters, *i.e.*, two sets of 2D coordinates on a pair of fixed parallel planes respectively which define NA's optical axis, and a remaining parameter that defines the distance of the NA camera to the eyeball. Finally, we use a two-stage iterative searching process to find the best NA camera pose parameters for each

image pair which achieves the best alignment quality.

Moreover, we set up a registration pipeline based on deep Convolutional Neural Networks (CNNs) which can make more robust prediction by learning from data. The pipeline consists of two vessel segmentation networks which extract vessel structures from both modalities respectively, a feature detection and description network to find correspondence between the two images, and an ourlier rejection network to filter out the outliers and keep the inliers among the correspondence. After training the networks, the UWF distortion correction model and the iterative searching algorithm are also integrated into the pipeline, to ensure that the UWF images have minimum distortions for feature extraction and alignment.

This paper extends our previous work [90] where deep CNNs and a distortion correction module with one global parameter are combined to align UWF with NA retinal images. In particular, the following extensions are considered in this paper: (1) We set up a more comprehensive correction function which is more capable of handling the complex distortion in UWF-to-NA alignment. Four more parameters are included to model the NA camera pose, so that the corrected UWF images will better match the NA image. (2) In conjunction with the new correction model, we set up an iterative searching process to find the best parameters. Moreover, we optimize the parameters for each image pair independently instead of finding a global parameter for the whole dataset, so that optimal results can be achieved for each case. (3) We collect a new dataset of UWF and NA images that contains more image pairs. Especially, we include more NA images covering the peripheral retinal areas to examine the effectiveness of our proposed distortion correction method.

The remaining of the paper is organized as follows. Section 3.2 provides backgrounds of UWF imaging and reviews related works of retinal image alignment. Section 3.3 sets up our proposed distortion correction functions. Section 3.4 presents the structures and training criteria of our alignment networks, and the searching algorithm of the distortion correction parameters in the testing phase. Section 3.5 provides experimental results of our methods.

3.2 Backgrounds

3.2.1 UWF Retinal Imaging

Optos UWF instrument follows the DICOM standard [16] in imaging and storing 3D retina data in 2D arrays with stereographic projection, as shown in Fig. 3.1. The eyeball is modeled as a pure sphere with its two poles located at the cornea and fovea respectively. As a special case of projective pinhole camera model, the UWF camera has its view point at cornea, its optical axis through cornea and fovea, and its imaging plane at the equator of the sphere.

The Optos UWF instrument uses an ellipsoidal mirror which has two focal points for this projection. It emits laser beams in varying angles at one of the focal points, which converge at the other focal point after being reflected by the mirror. Once the patients manage to locate their cornea at the second focal point, an UWF image can be captured.

3.2.2 Multi-Modal Retinal Image Alignment

Extensive works have been done for multi-modal retinal image registration. In feature-based alignment methods, multiple methods have been proposed to improve keypoint detection [25], feature description [26, 27], and matching and outlier rejection [28, 29]. The anatomical structures of retina can also help the alignment task, including vessels [24, 30–32], and vessel bifurcations and crossovers [24].

Recently, several CNN-based multi-modal retinal image registration algorithms have been proposed, which can be grouped into two categories. Some methods focus on estimating global transformation models, where CNN is usually adopted as alternative steps in feature-based registration pipelines. Lee *et al.* [34] proposed to filter extracted features based on their reliabilities predicted from their corresponding image patches with a CNN. Arikan *et al.* [35] proposed to use CNN for vessel segmentation and bifurcation detection. Wang *et al.* [85] set up fully CNN-based registration pipeline consisting of vessel segmentation, feature detection

and outlier rejection. Other methods mainly focus on deformable registration which is similar to optical flow estimation networks, whose capabilities are usually limited to aligning small displacements. The deformable alignment networks can be trained with weak supervision, *e.g.*, extracted vessels [39] from style transfer [18], monogenic phase signals [22], pixel gradients [38], or even synthesized images [33] via Generative Adversarial Networks [64]. Nevertheless, none of these methods is designated for the UWF-to-NA alignment task.

3.2.3 UWF / 3D Retinal Image Alignment

There exists very few works on image registration methods for UWF images. Ding *et al.* [36] proposed to align two retinal modalities of Optos UWF ColorMap and Fluorescein Angiography (FA) by 2nd order polynomial transformation model whose parameters are jointly optimized with a vessel segmentation network. However, there is no special consideration on the effects of UWF's peripheral distortions. Recently, they also proposed a two-stage registration method to align UWF FA image pairs [91]. In the first stage, an improved RANSAC method repeatedly estimates transformation matrices from the random parts of a high-confidence correspondence set, and then verify them on another noisier correspondence set. The matrix that achieves the most number of inliers during verification is considered optimal. Then, in the second stage, local chamfer alignment is adopted to reduce misalignments.

Nevertheless, several NA-to-NA retinal image registration methods incorporate 3D information to provide guidance to handle the UWF distortion. Hernandez-Matas *et al.* [87] proposed to align NA retinal image pairs via camera pose estimation with the assumption of spherical eyeball shape. Their objective function aims to minimize the errors between the back-projected 3D keypoint correspondence via ray tracing from both 2D images, which is optimized by Particle Swarm Optimization. Recently, they also adopt the assumption of ellipsoidal eyeball shape, whose parameters are jointly optimized with the extrinsic camera matrix [89]. Ataer-Cansizoglu *et al.* [88] also proposed to reconstruct 3D retina from multiple NA image via pose estimation

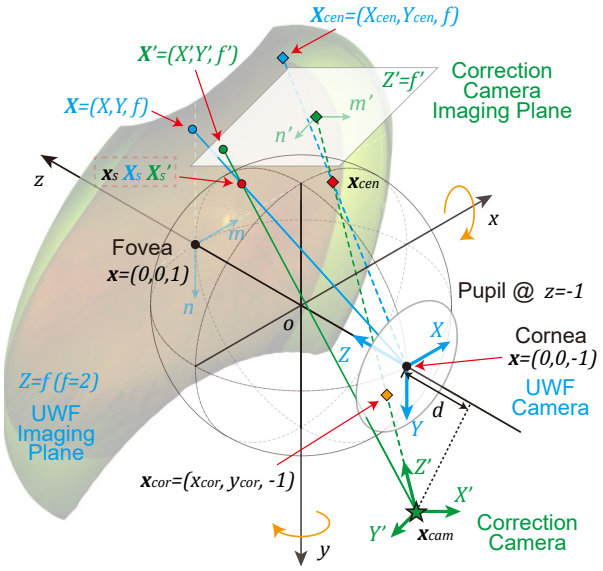


Figure 3.3: Illustration of the ideal UWF imaging process with a spherical eyeball shape assumption, as well as the correction camera model for perspective distortion correction process. Black, blue and green symbols represent coordinate systems for the world (xyz), UWF camera (XYZ) and correction camera ($X'Y'Z'$) respectively. In the world coordinate, the UWF camera is located at $(0,0,-1)$ and the correction camera is at $\mathbf{x}_{cam}$. A same point can be represented as $\mathbf{x}$ in the world coordinate, $\mathbf{X}$ in the UWF camera coordinate, or $\mathbf{X}'$ in the correction camera coordinate. (m,n) and (m',n') are 2D pixel coordinate systems for the UWF and corrected images respectively.

with spherical or quadratic eyeball shapes. They optimize the camera matrices and the shape parameters via bundle adjustment, which minimizes the overall 2D reprojection errors among all image pairs.

3.3 Proposed Distortion Correction in Ultra Widefield Images

3.3.1 Setups for UWF Imaging

We make the assumption of spherical eyeball. Based on UWF's imaging procedure, we set up a 3D coordinate system (*i.e.*, world coordinate) as shown in Fig. 3.3. We set the radius of sphere as $r = 1$ and the sphere center, fovea and cornea at the origin $(0,0,0)$, $(0,0,1)$ and $(0,0,\text{-}1)$ respectively. The UWF camera is placed at cornea with its direction towards -z and

its imaging plane perpendicular to the z axis. So the conversion between the UWF camera and world coordinate only involves a translation of $[0,0,\text{-}1]^T$. For simplicity, we also assume that the principle point of the pixel coordinate overlaps with the image plane center in each camera, and there is no skewing between the them.

In the UWF camera, given a position (m,n), $m,n \in \mathbb{R}$ in the UWF pixel coordinate, its camera coordinate (blue circle) can be written as

$$\mathbf{X} = [X,Y,f]^T = [m/\gamma, n/\gamma, f]^T, \tag{3.1}$$

where f is focal length and γ is a scaling factor between the pixel and image coordinates. Its corresponding 3D position on the eyeball (red circle) is the intersection of projection ray

$$\mathbf{X}_s = t\mathbf{X} \tag{3.2}$$

and the sphere

$$\|\mathbf{X}_s - \mathbf{O}_u\|^2 = 1 \tag{3.3}$$

where $\mathbf{O}_u = [0,0,1]^T$ is the eyeball center in UWF camera coordinate. t is solved as

$$t = \frac{2f}{X^2+Y^2+f^2} \tag{3.4}$$

where another solution $t = 0$ is the camera origin and thus not used. Therefore, its corresponding 3D point $\mathbf{x}_s$ in the world coordinate is

$$\mathbf{x}_s = \mathrm{F}_{sg}^{\text{-}1}(\mathbf{X}) = t\mathbf{X} + [0,0,\text{-}1]^T = \frac{(2fX, 2fY, f^2\text{-}X^2\text{-}Y^2)}{f^2+X^2+Y^2}. \tag{3.5}$$

On the other hand, a 3D point $\mathbf{x} = [x,y,z]^T$ in the world coordinate can be converted into the UWF camera coordinate $\mathbf{X} \leftarrow [x,y,z+1]^T$, and then projected onto the UWF imaging plane

as

$$\mathbf{X} = \mathrm{F}_{sg}(\mathbf{x}) = \left(\frac{fx}{1+z}, \frac{fy}{1+z}, f \right). \tag{3.6}$$

In this work, we set $f = 2$, *i.e.*, the UWF imaging plane is located at the back of the eyeball.

3.3.2 Correction Camera

In order to reduce UWF image's peripheral distortions during alignment, we set up a new correction camera (green star) that shares the same extrinsic parameters as the NA camera. The corrected UWF image from the correction camera will bear similar distortions as the NA image, which will improve the alignment quality. We use the following five parameters (in three sets) to define the correction/NA camera pose, which are illustrated in Fig. 3.3.

(1) We assume that the NA's image center is at $\mathbf{X}_{cen} = [X_{cen}, Y_{cen}, f]^T$ (blue diamond) after being aligned with the UWF modality, since both cameras might centered on different retinal areas. So, its corresponding 3D point on the eyeball (red diamond) is $\mathbf{x}_{cen} = \mathrm{F}_{sg}^{-1}(\mathbf{X}_{cen})$.

(2) The camera's optical axis (green dashed line) goes through the $\mathbf{x}_{cen}$ as well as $\mathbf{x}_{cor} = [x_{cor}, y_{cor}, -1]^T$ (orange diamond), because the pupil's aperture allows more flexible adjustment of the optical axis.

(3) We define $d > 0$ as the distance on the z axis from the correction camera's origin (green star) to the pupil's plane $z = -1$, since the NA camera looks from a more distant position than the UWF camera.

Therefore, the orientation of the correction camera is by $\vec{\mathbf{n}} = [\vec{n}_x, \vec{n}_y, \vec{n}_z]^T = \mathrm{normalize}(\mathbf{x}_{cen} - \mathbf{x}_{cor})$, and the camera origin is at

$$\mathbf{x}_{cam} = \mathbf{x}_{cor} - \frac{d}{\vec{n}_z}\vec{\mathbf{n}}. \tag{3.7}$$

We also set up rigid transformation (*i.e.*, rotation and translation) between the world coordinate and the correction camera coordinator. Based on right-hand rule, $\mathbf{R}_x$ is denoted as a rotation matrix which rotates an object point $\mathbf{x}$ around x axis by $\theta_x = \text{-}\tan^{\text{-}1}(\vec{n}_y/\sqrt{\vec{n}_z^2+\vec{n}_x^2})$, and $\mathbf{R}_y$ as the rotation around y by $\theta_y = \tan^{\text{-}1}(\vec{n}_x/\vec{n}_z)$. Therefore, $\mathbf{R} = \mathbf{R}_y\mathbf{R}_x$ represents the rotation between the two coordinates, while the translation between them is represented by $\mathbf{x}_{cam}$.

We set the correction camera's focal length as $f' = f + \|\mathbf{x}_{cam}\text{-}\mathbf{x}_{cor}\|$ and its pixel-to-image scaling factor as $\gamma' = \gamma$. Therefore, the UWF image's center area will not change in scale after correction with $\mathbf{X}_{cen} = [0,0,f]^T$ and $\mathbf{x}_{cor} = [0,0,\text{-}1]^T$.

3.3.3 Keypoint Remapping

Given a point $\mathbf{X} = [X,Y,f]^T$ (blue circle) on the UWF imaging plane, we would like to find its projected position $\mathbf{X}' = [X',Y',f']^T$ (green circle) in the correction camera coordinate. The corresponding 3D position of $\mathbf{X}$ on the eyeball (red circle) is $\mathbf{x}_s = \text{F}_{sg}^{\text{-}1}(\mathbf{X})$. Then we derive its coordinate in the correction camera as

$$\mathbf{X}'_s = [X'_s, Y'_s, Z'_s]^T = \mathbf{R}^{\text{-}1}(\mathbf{x}_s - \mathbf{x}_{cam}). \tag{3.8}$$

Its position on the correction camera's imaging plane is

$$\mathbf{X}' \leftarrow \frac{f'}{Z'_s}\mathbf{X}'_s. \tag{3.9}$$

This process computes the keypoint (m',n'), $m',n' \in \mathbb{R}$ in the corrected image which correspond to a UWF keypoint (m,n) detected in the original image. The complete process is summarized in Algorithm 1.

Algorithm 1: UWF keypoint correction from the correction camera view point.

Inputs: (m,n) (*e.g.*, UWF's keypoints), given 5 parameters $(\mathbf{X}_{cen}, \mathbf{x}_{cor}, d)$;
Outputs: (m',n') (*e.g.*, corrected keypoints);
(0) Compute correction camera pose x_{cam} and $\mathbf{R}$;
(1) Switch to UWF camera coordinate: $\mathbf{X} \leftarrow [m/\gamma, n/\gamma, f]^T$;
(2) Obtain 3D point in world coordinate: $\mathbf{x}_s \leftarrow \mathrm{F}_{sg}^{-1}(\mathbf{X})$;
(3) Switch to correction camera coordinate: $\mathbf{X}'_s \leftarrow \mathbf{R}^{-1}(\mathbf{x}_s - \mathbf{x}_{cam})$;
(4) Correction camera projection: $[X',Y',f']^T \leftarrow f'\mathbf{X}'_s/Z'_s$;
(5) Switch to pixel coordinate: $(m',n') \leftarrow \gamma'(X',Y')$.

3.3.4 Image Distortion Correction

On the other hand, when given a point in the correction camera $\mathbf{X}' = [X',Y',f']^T = [m'/\gamma', n'/\gamma', f']^T$ where (m',n') is its position in the correction camera pixel coordinate, we want to find its corresponding position $\mathbf{X} = [X,Y,f]^T$ in the UWF camera. We convert $\mathbf{X}'$ into the world coordinate as

$$\mathbf{x} = \mathbf{R}\mathbf{X}' + \mathbf{x}_{cam}. \tag{3.10}$$

The corresponding 3D point $\mathbf{x}_s$ is located at the intersection of the correction camera's projection ray (in world coordinate)

$$\mathbf{x}_s = \mathbf{x} + t \cdot (\mathbf{x}_{cam} - \mathbf{x}). \tag{3.11}$$

and the unit sphere

$$\|\mathbf{x}_s\|^2 = 1, \tag{3.12}$$

which leads to a 2nd-order equation

$$\|\mathbf{x}_{cam} - \mathbf{x}\|^2 t^2 + 2\mathbf{x}^T(\mathbf{x}_{cam} - \mathbf{x})t + \|\mathbf{x}\|^2 - 1 = 0. \tag{3.13}$$

Algorithm 2: Image generation from the correction camera view point (UWF image correction).

Inputs: (m'_p, n'_p) (pixel grid of the correction image), and the uncorrected UWF image I_{uwf}, given 5 parameters $(\mathbf{X}_{cen}, \mathbf{x}_{cor}, d)$;
Outputs: (m,n) (*e.g.*, resampling positions in I_{uwf}), corrected image I'_{uwf};
(0) Compute correction camera pose x_{cam} and $\mathbf{R}$;
(1) Switch to correction camera coordinate: $\mathbf{X}' \leftarrow [m'_p/\gamma', n'_p/\gamma', f']^T$;
(2) Switch to world coordinate: $\mathbf{x} \leftarrow \mathbf{R}\mathbf{X}' + \mathbf{x}_{cam}$;
(3) Get 3D intersection position: $\mathbf{x}_s \leftarrow \mathbf{x} + t \cdot (\mathbf{x}_{cam} - \mathbf{x})$, where t is got from Eq. (3.14);
(4) Switch to UWF coordinate and projection: $\mathbf{X} \leftarrow \mathrm{F}_{sg}(\mathbf{x}_s)$;
(5) Switch to pixel coordinate: $(m,n) \leftarrow \gamma(X,Y)$;
(6) Interpolating UWF image: $I'_{uwf} \leftarrow \mathrm{STN}(I_{uwf}, (m,n) - (m'_p, n'_p))$.

We only keep the smaller t when there are real value solutions, since we only need the intersection point closer to the back of the eyeball, *i.e.*,

$$t = \frac{-b - \sqrt{b^2 - 4ac}}{2a}, \tag{3.14}$$

where $a = \|\mathbf{x}_{cam}\text{-}\mathbf{x}\|^2$, $b = 2\mathbf{x}^T(\mathbf{x}_{cam}\text{-}\mathbf{x})$, and $c = \|\mathbf{x}\|^2$-1. Finally, the corresponding position in the UWF imaging plane can be obtained as $\mathbf{X} = \mathrm{F}_{sg}(\mathbf{x}_s)$.

This inverse mapping process is used to generate the corrected image by interpolation. We set (m'_p, n'_p), $m'_p, n'_p \in \mathbb{Z}$ as pixel grid positions in the correction image, and find their sampling positions (m,n) in the original UWF image. The complete process is listed in Algorithm 2.

3.3.5 Pixel Scaling Factor

The scaling factors between pixel and image/camera coordinates γ_x and γ_y are derived through the UWF center pixel view angles α and β in the DICOM files of Optos UWF images. As illustrated in Fig. 3.4, the width of the center pixel is $2f\tan(\alpha/4)$ in the X axis of UWF camera coordinate, which corresponds to width of 1 in the m axis of the pixel coordinate. Therefore, we set $\gamma_x = 1/\big(2f\tan(\alpha/4)\big)$. Optos UWF ColorMap has $\alpha = \beta = 0.08596515°$, so we set a same

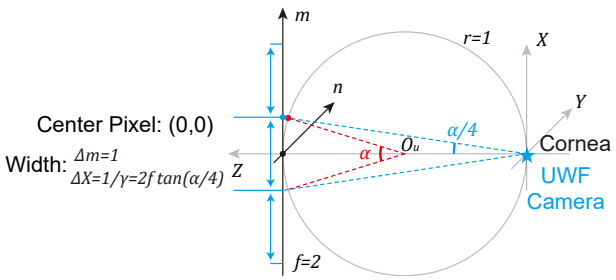


Figure 3.4: Illustration of scaling conversion between UWF pixel coordinate mn and camera coordinate XYZ. It only shows the projection on $X-Z$ plane and m axis, where α is the view angle of UWF image's center pixel in X direction.

scaling factor for both X and Y directions as $\gamma = \gamma_y = \gamma_x$, and we set $\gamma' = \gamma$ as described in Section 3.3.2.

3.4 Ultra Widefield and Narrow Angle Image Alignment

The design of our registration pipeline mainly follows the global alignment networks in our previously proposed multi-modal retinal image registration networks [85], which consists of three parts. First, both UWF and NA retinal images are fed into two vessel segmentation networks respectively to extract vessel maps. Then, a feature detection and description network finds keypoints from the vessel maps, and obtains their corresponding features as well. The keypoints from both images are matched with each other based on their feature similarities to obtain a set of correspondences. Finally, an outlier rejection network predicts inliers among the correspondences, based on which the transformation matrix can be estimated.

Furthermore, our proposed registration pipeline is differentiated from the previous work in the following aspects. On one hand, we adopt a segmentation network [36] specially pre-trained on the UWF modality which will ensure better segmentation quality, and propose to train the segmentation network for NA modality through a joint segmentation and deformable alignment scheme. On the other hand, our proposed distortion correction function in Algorithm 2 is incorporated into the pipeline, and the five extrinsic parameters for the correction camera model

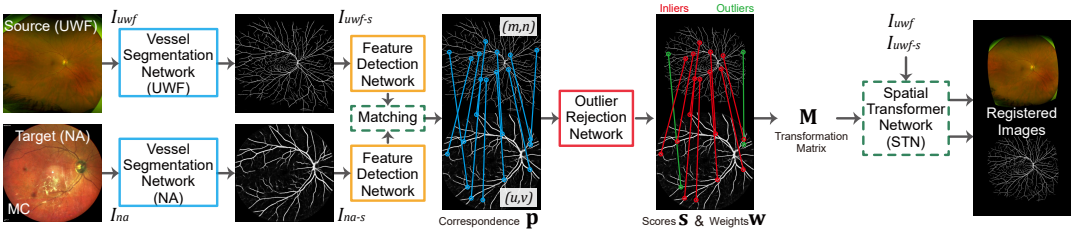


Figure 3.5: Our proposed registration pipeline: the networks.

are estimated through iterative alignment algorithms with the goal to achieve the best registration accuracy.

3.4.1 Vessel Segmentation Network

For the UWF modality, a pre-trained vessel segmentation network [36] is adopted to extract UWF vessel maps, because it has shown better segmentation performance on UWF images than our previously proposed network. Moreover, the knowledge stored in the UWF segmentation network can be migrated into the network for NA vessel segmentation, since both images are captured from the same eye and therefore should share a similar vessel structure. However, due to the severe distortions, the vessel structures are not accurately aligned even after global transformations, which will affect the efficiency in transferring vessel knowledge. In order to achieve this goal, we set up a training method which trains the NA's vessel segmentation network along with a deformable alignment network. During training, the predicted NA vessel map should be able to align with the UWF vessel map extracted from the pre-trained network, such that the vessel knowledge can be accurately transferred from the UWF segmentation network into the NA's network.

The detailed training scheme is shown in Fig. 3.6. Both NA image I_{na} and UWF images I_{uwf} are sent into their corresponding vessel segmentation networks to obtain vessel maps, $I_{uwf\text{-}s}$ and $I_{na\text{-}s}$, respectively. Meanwhile, both images are also concatenated and fed into the deformable alignment network to obtain a dense registration field F, which is similar to optical

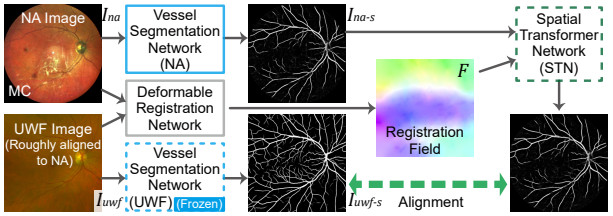


Figure 3.6: Training scheme for NA vessel segmentation network.

flow estimation [19]. Afterwards, a Spatial Transformation Network STN [58] (*i.e.*, an image warper which can back-propagate) warps the NA's vessel map according to the registration field. Finally, a loss function examines the alignment quality between the warped NA and UWF vessel maps, which helps transfer the UWF's vessel knowledge (*i.e.*, "ground-truth") into the NA vessel segmentation task.

The training loss is designed as

$$L_{seg} = \lambda_{pc} L_{pc} + \lambda_{sm} L_{sm} + \lambda_{rot} L_{rot} \tag{3.15}$$

where L_{pc} and L_{sm} are photometric consistency and smoothness losses in optical flow estimation networks [19], L_{rot} is rotation invariant loss [86], and λ_{pc} and λ_{sm} are their corresponding weights. L_{pc} is defined as

$$L_{pc}(I_{uwf\text{-}s}, I_{na\text{-}s}, F) = \text{MSE}\big(\text{STN}(I_{na\text{-}s}, F), I_{uwf\text{-}s}\big) \tag{3.16}$$

where MSE (Mean Square Error) aims to minimize the difference of the vessel segmentation results from both networks. Besides, L_{sm} is defined as

$$L_{sm}(F) = \underset{c,x,y}{\text{mean}}\left((F_{c,x,y} - F_{c,x+1,y})^2\right) + \underset{c,x,y}{\text{mean}}\left((F_{c,x,y} - F_{c,x,y+1})^2\right) \tag{3.17}$$

where F has two channels (*i.e.*, $c \in \{0,1\}$) of maps with the same spatial resolution as the NA image, and c, x and y are indices in F for the channel, horizontal, and vertical positions respectively. It ensures that neighboring pixels are warped in similar directions and distances

when the NA image is being warped, so that non-vessel areas can also be aligned correctly. L_{rot} is defined as

$$L_{rot}(I_{na}) = \text{MSE}\Big(\text{rot}\Big(\text{H}\big(\text{rot}(I_{na})\big)\Big), \text{H}(I_{na})\Big) \tag{3.18}$$

where $\text{rot}(I)$ rotates its input image by 180°, to ensure that the extracted vessels match the actual vessel positions.

For NA vessel segmentation and deformable alignment, we adopt the same network structures as those in our previously proposed fine alignment network [86], *i.e.*, Deep Retinal Image Understanding (DRIU) network [69] for NA vessel segmentation, and a modified structure based on U-Net [65] for deformable alignment. Readers can refer to [86] for detailed network structures. These networks are trained on roughly aligned NA and UWF images, since the deformable alignment network could not handle large displacements. The UWF images for training are warped by affine matrices estimated from manually labeled keypoints. After training, only the vessel segmentation networks are used in the following steps, while the deformable alignment network is not used.

3.4.2 Feature Detection and Description

In this paper, SuperPoint [20] is adopted to extract features from both vessel segmentation maps. The network consists of an encoder and two decoders for keypoint prediction and feature description respectively. The encoder first takes as input a vessel segmentation map ($I_{uwf\text{-}s}$ or $I_{na\text{-}s}$), whose output is sent into both decoders. Afterwards, the keypoint decoder outputs a keypoint probability map where keypoints are then located through Non-Maximun Suppression. The keypoint coordinates for UWF and NA images are denoted as (m,n) and (u,v) respectively. In parallel, the feature description decoder outputs a feature tensor, where the feature descriptor at each keypoint can be extracted.

Two loss functions, *i.e.*, keypoint loss and descriptor loss, are involved in training the

network. The keypoint loss is a cross-entropy loss to supervise the multi-class classification of keypoint and non-keypoints in a small image area. While the descriptor loss minimizes the feature differences at the corresponding positions between synthetically warped image pairs. In this paper, we adopt a pre-trained SuperPoint model in prediction, because complete keypoint positions and accurate transformation matrices for our data are not available. The model was first trained with both loss functions on a synthetic dataset [20] which consists of synthesized images of corners and edges as well as their keypoint position ground-truths. Furthermore, the feature description decoder was finetuned with only the descriptor loss on retinal vessel maps extracted from a NA-to-NA alignment dataset [85]. Readers can refer to [20, 85] for more details on the training data and process.

Following the SuperPoint network, a bi-directional matching process finds initial correspondence $\mathbf{p} = [..., (m_i, n_i, u_i, v_i), ...] \in \mathbb{R}^{N\times 4}$ by matching the features from both images. More specifically, in (m_i, n_i, u_i, v_i), the UWF feature at (m_i, n_i) should be a best match for the NA feature at (u_i, v_i) among all UWF features based on minimum euclidean distance, and vice versa.

3.4.3 Outlier Rejection Network

The initial correspondence might be noisy due to the large view angle differences of the two modalities. The UWF's features in peripheral areas could be matched incorrectly with the NA's features, which will affect the accuracies of the estimated transformation matrices. Therefore, an outlier rejection network [47] is adopted to predict the inliers and outliers among all correspondences, and the transformation parameters are estimated based on the inliers.

The outlier rejection network takes $\mathbf{p}$ as its input, and outputs scores $\mathbf{s} \in \mathbb{R}^{N\times 1}$ for every correspondence. Then, the scores are thresholded by 0, and transformed into weights $\mathbf{w}$ in range $[0, 1)$, *i.e.*, $\mathbf{w} = tanh(ReLU(\mathbf{s}))$ The keypoint pairs with $\mathbf{w}_i > 0$ are inliers, while those with $\mathbf{w}_i = 0$ are considered as outliers which will have no effect on the following steps. Afterwards, the matrices $\mathbf{M}_t$ of most transformations can be estimated by Weighted Least Square (WLS)

algorithm from **p** and **w**. For a transformation

$$(u_i, v_i) = T\big((m_i, n_i), \mathbf{M}_t\big), \tag{3.19}$$

the matrix can be solved by

$$\mathbf{M}_t = \arg\min_{\mathbf{M}} ||\mathbf{W}\mathbf{A}_t \mathrm{Vec}(\mathbf{M})||, \tag{3.20}$$

where $t \in \{poly, pers, DLT\}$ represents transformation types, $\mathbf{W} = \mathrm{diag}([w_1, w_1, ..., w_N, w_N]) \in \mathbb{R}^{2N\times 2N}$ is a diagonal matrix, $\mathrm{Vec}(\mathbf{M})$ is vectorized $\mathbf{M}$, and $\mathbf{A}_t$ is constructed according to t. The solution to Eq. (3.20) is the eigenvector that corresponds to the smallest eigenvalue of $\mathbf{A}^T\mathbf{W}^T\mathbf{W}\mathbf{A}$.

For 2nd-order polynomial transformation ($t = poly$)

$$[u_i, v_i]^T = \mathbf{M}_{poly}[m_i, n_i, m_i^2, n_i^2, m_i n_i, 1]^T, \tag{3.21}$$

where

$$\mathbf{M}_{poly} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} \end{pmatrix}, \tag{3.22}$$

the matrix $\mathbf{A}_{poly} \in \mathbb{R}^{2N\times 13}$ is constructed as

$$\mathbf{A}_{poly} = \begin{pmatrix} \dots & & & & & & & & & & & & \dots \\ m_i & n_i & m_i^2 & n_i^2 & m_i n_i & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \text{-}u_i \\ 0 & 0 & 0 & 0 & 0 & 0 & m_i & n_i & m_i^2 & n_i^2 & m_i n_i & 1 & \text{-}v_i \\ \dots & & & & & & & & & & & & \dots \end{pmatrix}. \tag{3.23}$$

For perspective transformation ($t = pers$)

$$\begin{pmatrix} u_i q \\ v_i q \\ q \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} m_i \\ n_i \\ 1 \end{pmatrix}, \tag{3.24}$$

the matrix $\mathbf{A}_{pers} \in \mathbb{R}^{2N \times 9}$ is constructed as

$$\mathbf{A}_{pers} = \begin{pmatrix} \dots & & & & & & & & \dots \\ m_i & n_i & 1 & 0 & 0 & 0 & -m_i u_i & -n_i u_i & -u_i \\ 0 & 0 & 0 & m_i & n_i & 1 & -m_i v_i & -n_i v_i & -v_i \\ \dots & & & & & & & & \dots \end{pmatrix}. \tag{3.25}$$

For Direct Linear Transformation (DLT)

$$\begin{pmatrix} u_i q \\ v_i q \\ q \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix} \begin{pmatrix} x_i \\ y_i \\ z_i \\ 1 \end{pmatrix}, \tag{3.26}$$

where (x_i, y_i, z_i) is the 3D world coordinate of the UWF keypoint (m_i, n_i) calculated by Eq. (3.1) and (3.5), the matrix $\mathbf{A}_{DLT} \in \mathbb{R}^{2N \times 12}$ is constructed as

$$\mathbf{A}_{DLT} = \begin{pmatrix} \dots & & & & & & & & & & & \dots \\ x_i & y_i & z_i & 1 & 0 & 0 & 0 & 0 & -x_i u_i & -y_i u_i & -z_i u_i & -u_i \\ 0 & 0 & 0 & 0 & x_i & y_i & z_i & 1 & -x_i v_i & -y_i v_i & -z_i v_i & -v_i \\ \dots & & & & & & & & & & & \dots \end{pmatrix}. \tag{3.27}$$

While for affine transformation $\mathbf{M}_{aff}$, readers can refer to [86] for its solution.

The loss function for network training consists of three parts, *i.e.*, classification loss L_c and regression loss L_r which are based on ground-truths matrix $\mathbf{M}_{gt}$, as well as an unsupervised *Dice* loss L_d based on alignment quality of the two vessel maps. First, L_c is defined as

$$L_c(\mathbf{p},\mathbf{s},\mathbf{M}_{gt}) = \frac{1}{N}\sum_{i=1}^{N}\gamma_i \mathrm{BCE}\big(y_i,\sigma(s_i)\big) \tag{3.28}$$

where $\mathrm{BCE}(\cdot,\cdot)$ is Binary Cross Entropy function, $\sigma(\cdot)$ is a sigmoid function, and γ_i is a balancing weight for two classes. $y_i \in \{0,1\}$ is the inlier label for the i-th correspondence

$$y_i = \begin{cases} 1, & \big|\big|T\big((m_i,n_i),\mathbf{M}_{gt}\big) - (u_i,v_i)\big|\big| \leq 5 \text{ pixels} \\ 0, & \text{otherwise} \end{cases} \tag{3.29}$$

where $T(\cdot)$ translates the UWF keypoint (m_i,n_i) into NA's coordinate based on $\mathbf{M}_{gt}$, *i.e.*, a keypoint pair is labeled as an inlier if their distance is less than 5 pixels after warping. Next, L_r aims to minimize the difference between $\mathbf{M}_{gt}$ and the estimated $\mathbf{M}_{aff}$, *i.e.*,

$$L_r = \mathrm{MSE}(\mathbf{M}_{gt} - \mathbf{M}_{aff}). \tag{3.30}$$

Finally, L_d aims to further improve the alignment quality by comparing the vessel maps of NA and UWF after warping, which is useful when the ground-truths labels lack accuracies [86] or the supervised loss values are enlarged by the huge non-linear distortions. L_d is defined as

$$L_d = 1 - Dice\Big(\mathrm{STN}(I_{uwf\text{-}s},\mathbf{M}_{poly}), I_{na\text{-}s}\Big), \tag{3.31}$$

where *Dice* is an evaluation metric on image alignment quality

$$Dice(I_1,I_2) = \frac{2\cdot\sum\big(\mathrm{ele\_min}(I_1,I_2)\big)}{\sum I_1 + \sum I_2}. \tag{3.32}$$

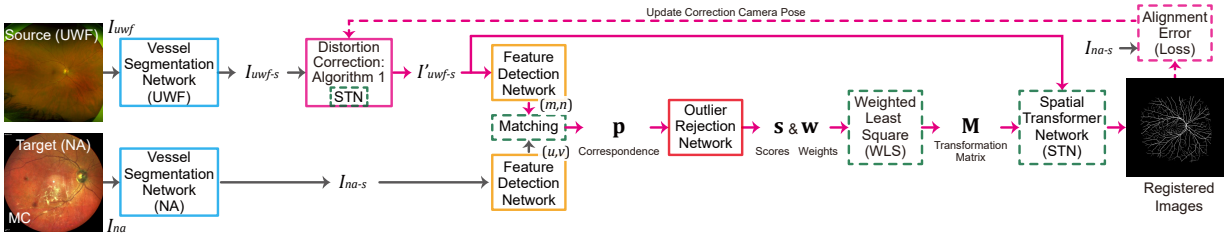


Figure 3.7: Our proposed registration pipeline: the iterative searching process for distortion correction. The magenta arrows indicates the iterative searching loop, where the solid arrows represent the networks' computations with the corrected I'_{uwf} and the dashed arrows indicate alignment error computation and camera pose update.

Therefore, the total loss for the outlier rejection network is

$$L_{outlier} = \lambda_c L_c + \lambda_r L_r + \lambda_d L_d. \tag{3.33}$$

Readers can also refer to [85] for more details. Different from [85], we use $\mathbf{M}_{poly}$ for L_d because higher order polynomial transformation can better describe the non-linear distortions in our task, which is aimed to reduce the ground-truths' misalignment errors during training. Meanwhile we still use affine transformation for $\mathbf{M}_{gt}$, L_c and L_r due to its stability.

3.4.4 Iterative Alignment Algorithm

Even with the aforementioned learnable methods, the UWF-to-NA alignment performance is still affected by the perspective distortion in two aspects. On one hand, the adopted 2D-to-2D global transformation models may not be able to handle the nonlinear distortions stemming from the sphere-to-plane projection, especially when aligning NA images with the UWF's peripheral areas. On the other hand, in the network forward paths, the keypoint features are extracted from the heavily distorted UWF images and the less distorted NA images. As a result, the feature descriptors on a same vessel structure from the two images will bear larger variations, leading to more errors in the following matching and outlier rejection process.

In order to further reduce alignment error, we incorporate the aforementioned distortion

Algorithm 3: Global searching for initial d.

Result: Optimal $d_{init} = d$.
Initialization:
(1) $\{(I^{(j)}_{uwf\text{-}s}, I^{(j)}_{na\text{-}s}), ...\}$ of whole training set, and NA's keypoints $\{[(u^{(j)}_i, v^{(j)}_i)], ...\}$;
(2) $d = 1, \mathbf{X}_{cen} = (0, 0, f), \mathbf{x}_{cor} = (0, 0, \text{-}1)$;
(3) A dictionary D to store results;
(4) Searching range $[be, en]$, $step$, ext, and $loops$;
for $loops$ **do**
  for $d_c = (dstep * 2) : (step) : (d + step * 2)$ **do**
    for j-th image pair **do**
      Algorithm 2: $I^{(j)}_{uwf\text{-}s} \rightarrow I^{(j)'}_{uwf\text{-}s}$;
      Feature Detection & Description Network: $I^{(j)'}_{uwf\text{-}s} \rightarrow [(m^{(j)'}_i, n^{(j)'}_i)]$;
      Outlier Rejection Network & WLS: $[(m^{(j)'}_i, n^{(j)'}_i, u^{(j)}_i, v^{(j)}_i)] \rightarrow \mathbf{M}^{(j)}$;
      $D[d_c][j] = Dice(\text{STN}(I^{(j)'}_{uwf\text{-}s}, \mathbf{M}^{(j)}), I^{(j)}_{na\text{-}s})$
    end
    $D[d_c] = \text{Average}_j(D[d_c][j])$;
  end
  $d \leftarrow \arg\max_{d_c} D[d_c]$;
  $step = step/2$;
end

correction process into our registration pipeline in the testing phase, such that the transformation models are estimated based on the corrected UWF images with less distortions. Meanwhile, we set up two algorithms to search for the optimal distortion correction parameters that yield the best alignment quality according to *Dice* scores. The searching process consists of two steps, *i.e.*, a global searching process on the whole training dataset to find an initial d_{init}, and a local searching process to find the optimal parameters on each image pair to be aligned.

Algorithm 3 lists the first global searching process for d_{init}, which remains the same as our previous work [90]. Here, we set $\mathbf{X}_{cen} = (0, 0, f)$ and $\mathbf{x}_{cor} = (0, 0, \text{-}1)$, *i.e.*, we assume the average optical axis of all training NA images overlaps with the z axis. For each d candidate to be evaluated, Algorithm 2 first remaps all UWF images in the training set based on current correction camera pose. Then, new UWF keypoints are detected by the feature detection and

Algorithm 4: Iterative local searching and alignment.

Result: Registered image I_{uwf}^{reg}.
Initialization:
(1) A set of I_{uwf}, I_{na}, $I_{uwf\text{-}s}$ and $I_{na\text{-}s}$ in test set;
(2) **p** and **w** from initial alignment;
(3) $\mathbf{X}_{cen} = (0,0,f)$, $\mathbf{x}_{cor} = (0,0,\text{-}1)$, $d = d_{init}$;
(4) A dictionary D to store *Dice* values.
Function `NetworkForward()`**:**
    Algorithm 2: $I_{uwf}, I_{uwf\text{-}s} \rightarrow I'_{uwf}, I'_{uwf\text{-}s}$;
    Feature Detection & Description Network: $I'_{uwf\text{-}s} \rightarrow [(m'_i, n'_i)]$;
    Outlier Rejection Network & WLS: $[(m'_i, n'_i, u_i, v_i)] \rightarrow \mathbf{M}$;
End Function
Obtain $\mathbf{X}_{cen}$ by NA-to-UWF transformation: Eq. (3.34);
while *True* **do**
    while *True* **do**
        for $d_c = (d - step*2) : (step) : (d + step*2)$ **do**
            `NetworkForward()` $\rightarrow I'_{uwf\text{-}s}, \mathbf{M}$;
            $D[d_c, \mathbf{X}_{cen}, \mathbf{x}_{cor}] = Dice(\text{STN}(I'_{uwf\text{-}s}, \mathbf{M}), I_{na\text{-}s})$
        end
        $d \leftarrow \arg\max_{d_c} D[d_c, ...]$;
        if d not updated in this loop **then**
            break;
        end
    end
    while *True* **do**
        for $\mathbf{x}_{corc} \in$ 8 neighbors around $\mathbf{x}_{cor}$ **do**
            `NetworkForward()` $\rightarrow I'_{uwf\text{-}s}, \mathbf{M}$;
            $D[d, \mathbf{X}_{cen}, \mathbf{x}_{cor\text{-}c}] = Dice(\text{STN}(I'_{uwf\text{-}s}, \mathbf{M}), I_{na\text{-}s})$
        end
        $\mathbf{x}_{cor} \leftarrow \arg\max_{\mathbf{x}_{cor\text{-}c}} D[..., \mathbf{x}_{cor\text{-}c}]$;
        if $\mathbf{x}_{cor}$ not updated in this loop **then**
            break;
        end
    end
    if both d & $\mathbf{x}_{cor}$ not updated in this loop **then**
        break;
    end
end
`NetworkForward()` $\rightarrow I'_{uwf}, \mathbf{M}$;
Get final image: $I_{uwf}^{reg} = \text{STN}(I'_{uwf}, \mathbf{M})$.

description network which are then matched with the NA images' keypoints as correspondences. Afterwards, the transformation matrices are updated, and the corrected UWF images are warped again based on their respective matrices which are compared with the NA images to compute *Dice* scores. The d with the highest average *Dice* value on the training set is selected as the d_{init} for the next step. The searching settings in Algorithm 3 are empirically set as $be = 1$, $en = 3$, $step = 1/2$, $ext = 4$, and $loop = 4$ at the beginning.

Then, in the second step, Algorithm 4 is proposed to find the locally optimal 5 parameters on each testing image pair individually. In brief, we first estimate the $\mathbf{X}_{cen}$, which is obtained by warping NA's center pixel coordinate $(u_0, v_0) = (0,0)$ to the UWF image space (*i.e.*, inverse transformation)

$$[m_{cen}, n_{cen}]^T = \mathbf{M}_{poly}^{-1}[u_0, v_0, u_0^2, v_0^2, u_0 v_0, 1]^T, \tag{3.34}$$

and then scaled into the UWF camera coordinate via Eq. (3.1). $\mathbf{M}_{poly}^{-1}$ is an inverse transformation which is estimated from uncorrected $\mathbf{p}$ and $\mathbf{w}$ through WLS by swapping $[u_i, v_i]$ and $[m_i, n_i]$ in Eq. (3.23). Then, we optimize d and $\mathbf{x}_{cor}$ sequentially and iteratively in a big loop as shown in Fig. 3.7, until none of the parameters are updated in a loop. Both the feature detection and description and outlier rejection networks are involved in this optimization process, so that they can benefit from the distortion-corrected UWF images in feature description, matching and inlier predictions. When searching for d, we set $step = 1/16$ which is the minimum step used in Algorithm 3. Besides, when optimizing $\mathbf{x}_{cor}$, we evaluate the 8-neighbors of the current best value, *i.e.*, $\Delta X_{cor}, \Delta Y_{cor} \in \{-step, 0, step\}$ where $step = 1/20$.

3.4.5 Alignment Process

We set up several alignment methods that use various transformation models or distortion correction algorithms, which are summarized in Fig. 3.8. On one hand, all networks are evaluated on the uncorrected images with perspective, polynomial and DLT transformations, which are

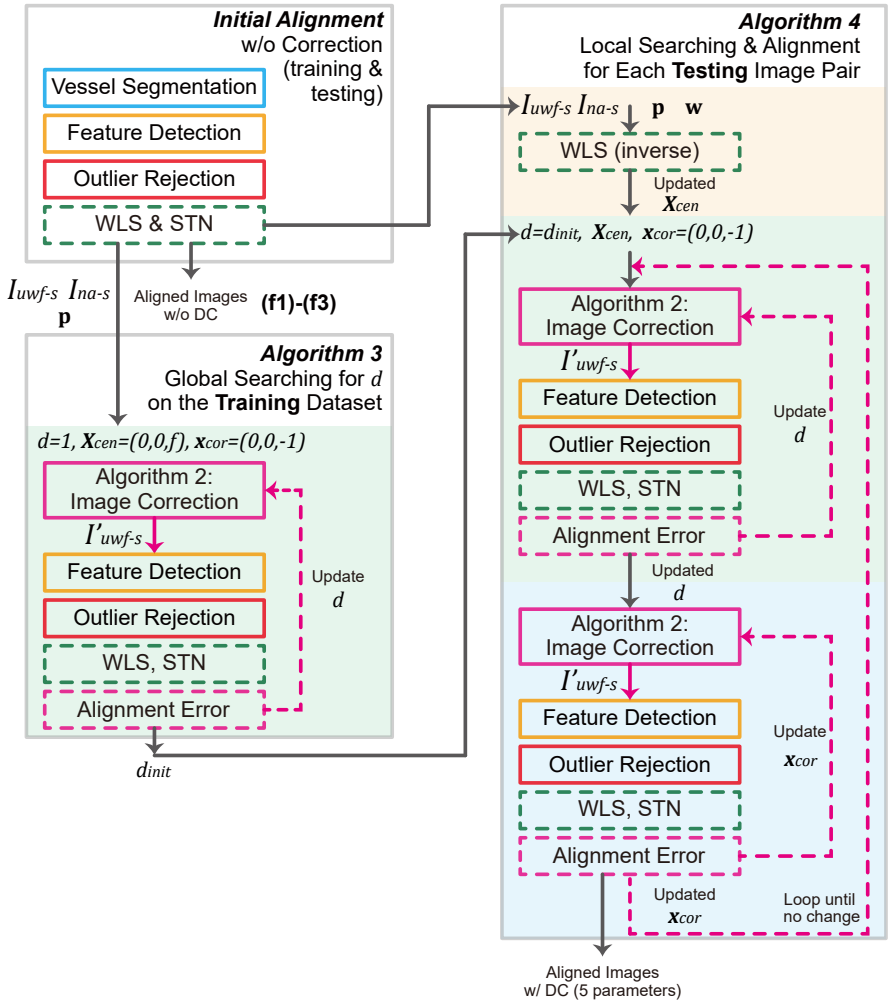


Figure 3.8: The complete registration and distortion correction process of our proposed method.

denoted as (f1)-(f3) respectively. On the other hand, we search for the best five parameters for each testing case individually by Algorithm 3 and 4, and corrects the UWF image for alignment with polynomial transformation, which is denoted as (f5) w/ DC (5 parameters).

In addition, in our ablation study, we set up alternative schemes with different network involvements in searching for best parameters in Algorithm 3 and 4. Scheme (f7) sends the corrected UWF keypoint positions into the outlier rejection network to re-estimate the weights for inliers, *i.e.*, feature detection and description network is not used in searching. It relies on the initial correspondence detected from the uncorrected vessel maps. Scheme (f6) only re-estimates

the transformation matrix with corrected UWF keypoints (by Algorithm 1) and uncorrected inlier weights without using any networks, whose alignment performance should suffer most from distortions.

3.5 Experiments

3.5.1 Dataset

We experiment on two datasets that consist of UWF and NA image pairs collected by Jacobs Retinal Center of University of California San Diego. In both datasets, the UWF images are stored in DICOM format in the Optos ColorMap with 4000×4000 resolutions, where $\alpha = \beta = 0.08596515^\circ$ for coordinate conversion are specified in the files. The NA images are captured in the 55° MultiColor (MC) modality from the Heidelburg Spetrallis system with image resolutions of 768×768 or 1536×1536. For convenience, we resize all MC images to 768×768 before alignment. In addition, we only use the center 2000×2000 areas of UWF images in our experiments, because only a few effective UWF pixels appear outside this region and most MC images fall in this area after alignment.

The first dataset, which we denote as UW, contains 116 pairs of images from 59 patients (*i.e.*, 2 patients with one eye, 57 patients with both eyes). The majority of the MC images locate around the fovea, *i.e.*, $\mathbf{X}_{cen}$ is close to $(0,0)$. We divide the dataset into two equal-sized sets, UW1 and UW2. We denote UW1-2 as the scheme of training on UW1 and testing on UW2, and vice versa.

Furthermore, we also collect a new dataset, *i.e.*, UWc, in this paper. The dataset contains images on 56 eyes from 29 patients (*i.e.*, 2 patients with one eye, 27 patients with both eyes). However, on each eye, the doctors captured multiple MC images centered on various retinal positions by manipulating the Spetrallis camera. As a result, there are 56 UWF images and 505 MC images, where each UWF image has 8 to 10 corresponding MC images. A plot that

summarizes the MC's center position after NA-to-UWF alignment based on ground-truths is shown in Fig. 3.10. We divide this dataset into UWc1 (250 pairs) and UWc2 (255 pairs). Similarly, we denote UWc1-2 as the scheme of training on UWc1 and testing on UWc2, and vice versa.

In order to provide ground-truths for training and testing, we manually labeled corresponding keypoint positions for all image pairs. The average numbers (standard deviation) of keypoint pairs among all images are 43.7 (18.6) for UW and 14.3 (4.2) for UWc. Depending on the image quality, we ensure to label at least 6 pairs of non-collinear correspondence for each image pair, which is verified by estimating polynomial transformation or Direct Linear Transformation (DLT) matrices.

3.5.2 Settings

In training the vessel segmentation networks for MC images, we initialize the networks with weights pre-trained on Color Fundus Images [86], and then finetune them on MC images in the UW1 or UW2 datasets. The networks are trained for 2000 epoches on both datasets, with learning rate as $1e$-3, batch size as 1, $\lambda_{sm} = 2e$-3, and $\lambda_{pc} = \lambda_{rot} = 1e$-3. For the outlier rejection networks, we also initialize the networks with a model pre-trained on a NA-to-NA retinal image alignment task [85], and then finetune them on UW1 or UW2 datasets respectively. The networks are trained for 1000 epoches, with learning rate as $1e$-4, batch size as 8, $\lambda_c = 1$, and $\lambda_r = \lambda_d = 0.1$. The models that achieve the highest average *Dice* scores on the training sets are used for evaluation. All networks are trained by Adam optimizer [72]. When evaluating our method on the UWc1 & UWc2 dataset, we directly use the networks trained on UW1, as well as the d_{init} estimated on UW1 as the input for Algorithm 4, *i.e.*, there is no network training or fine-tuning and Algorithm 3 is not executed.

In testing phase, 2nd-order polynomial transformation is adopted for all methods unless noted. We adopt two evaluation metrics for the alignment quality, *i.e.*, *Dice* of Eq. (2.8), and alignment success rate. For *Dice*, the vessel maps from the vessel segmentation networks in the

proposed pipeline are used in the computation. When computing success rates, the ground-truth keypoint coordinates in each UWF images are first corrected by Algorithm 1, and then warped based on their respective transformation matrices. Next, the pixel distances between the warped UWF keypoints and their corresponding NA keypoints are calculated, among which the maximum distance d_{max} is recorded. An image pair is considered as success in alignment if its d_{max} is smaller than a $threshold \in \{10, 20, 50\}$.

Our method is implemented in Python and PyTorch. The networks are trained on desktops and servers with GTX 1080 Ti. It takes about 18 hours to train a segmentation network and about 10 hours for the outlier rejection network. All the evaluations are performed on a Windows desktop with a GTX 1080 Ti and an Intel i7-7700K CPU.

3.5.3 Registration Results

For comparison, we set up 4 groups of registration methods in evaluation, as listed in Table 3.1. The methods are:

(a-b) Keypoint-matching based methods.

For (a1) Phase + HoG+ + RANSAC based on the global registration step in [23], dense HoG features are extracted from the local phase maps of both images, and then matched with each other. On UW datasets, we only use the center 1000×1000 area of UWF images as most NA images fall in this region after alignment. We do not report it on UWc dataset due to excessive memory cost in dense-HoG feature matching.

For (b1)-(b3), we set up the global registration method in [91]. SIFT [42] or SuperPoint features are detected on vessel segmentation maps. RANSAC and RANSAC-CS (RANSAC with Consensus and Sample sets) [91] are used to estimate inliers. In RANSAC-CS which is designed to align two UWF images, the optimal threshold for creating the Sample set is searched on UWc2 dataset at step 0.1 based on average *Dice* value, which is determined to be 0.7 and then applied

Table 3.1: Average *Dice* Values and Success Rates for UWF-to-NA Image Alignment

#	Method	Transformation	UW1-2 (58)				UW2-1 (58)				UWc1-2 (255)				UWc2-1 (250)			
			Dice	Success Rate			*Dice*	Success Rate			*Dice*	Success Rate			*Dice*	Success Rate		
				10	20	50		10	20	50		10	20	50		10	20	50
	Before Alignment		0.1721	0	0	0	0.1841	0	0	0	0.1649	0	0	0	0.1611	0	0	0
(a1)	Phase + HoG + RANSAC [23]	Affine	0.3555	/	/	/	0.3476	/	/	/	*[1]	*[1]	*[1]	*[1]	*[1]	*[1]	*[1]	*[1]
(b1)	SegMap + SuperPoint + RANSAC	Polynomial	0.4097	5.2	36.2	79.3	0.3626	1.7	24.1	63.8	0.4009	10.6	43.1	69.8	0.3382	7.2	26.0	50.0
(b2)	SegMap + SIFT + RANSAC-CS [91]	Polynomial	0.3985	3.4	37.9	72.4	0.3720	1.7	17.2	60.3	0.4162	9.4	38.0	70.4	0.3503	6.0	21.6	50.4
(b3)	SegMap + SuperPoint + RANSAC-CS [91]	Polynomial	0.4075	3.4	39.7	87.9	0.3775	5.2	25.9	81.0	0.4133	14.1	45.5	80.4	0.3701	8.6	34.5	63.5
(c1)	SegMap + SuperPoint + REMPE [89]	3D	0.2600	*[2]	*[2]	*[2]	0.2294	*[2]	*[2]	*[2]	0.2227	*[2]	*[2]	*[2]	0.1811	*[2]	*[2]	*[2]
(d1)	DRMIME [81]	Perspective	0.1687	0.0	0.0	0.0	0.1777	0.0	0.0	0.0	-	-	-	-	-	-	-	-
(d2)	SegMap + DRMIME [81]	Perspective	0.3540	3.4	34.5	48.3	0.3780	12.1	53.4	58.6	0.1971	0.0	5.5	6.3	0.1893	0.8	3.6	5.6
(e1)	DLIR [14]	Affine	0.1763	0.0	0.0	0.0	0.1878	0.0	0.0	0.0	0.1217	0.0	0.0	0.0	0.1152	0.0	0.0	0.0
(f1)	Our Network w/o DC	Perspective	0.4454	6.9	74.1	93.1	0.4409	13.8	79.3	96.6	0.3853	0.4	9.8	58.0	0.3574	0.8	10.0	58.0
(f2)	Our Network w/o DC	Polynomial	0.4973	20.7	67.2	93.1	0.4864	13.8	70.7	96.6	0.4882	11.8	49.8	93.3	0.4380	7.2	40.4	80.8
(f3)	Our Network w/o DC	DLT	0.4764	46.6	81.0	89.7	0.4515	37.9	74.1	89.7	0.4614	19.2	53.7	77.6	0.4020	22.8	45.6	60.4
(f4)	Our Network w/ DC (1 parameter) [90]	Polynomial	0.5415	44.8	87.9	93.1	0.5189	46.6	93.1	98.3	0.5131	34.1	69.0	92.5	0.4628	30.4	61.6	82.8
(f5)	Our Network w/ DC (5 parameters)	Polynomial	**0.5555**	44.8	91.4	93.1	**0.5331**	46.6	93.1	96.6	**0.5627**	54.1	84.7	97.6	**0.5184**	48.4	79.6	93.2

[1]Image size of UWF is too large for dense-HoG feature matching. [2]Cannot obtain accurate UWF keypoints after warping.

Table 3.2: Ablation Study of UWF-to-NA Image Alignment on UWc Datasets

#	Method	UWc1-2 (255)				UWc2-1 (250)				Running Time (Algorithm 4)		
		Dice	Success Rate			*Dice*	Success Rate			# Iters /pair	Time s/pair	Time s/iter
			10	20	50		10	20	50			
(f6)	Distortion Correction w/o Networks	0.5403	30.2	67.1	89.0	0.4914	24.7	56.0	77.6	131.68	65.30	0.48
(f7)	Distortion Correction before Outlier Rejection	0.5465	40.8	78.0	95.3	0.4917	37.6	67.6	84.8	57.84	35.25	0.60
(f5)	Distortion Correction before Feature Detection	0.5627	54.1	84.7	97.6	0.5184	48.4	79.6	93.2	23.44	22.55	0.96
(b1)	SegMap + SuperPoint + RANSAC	0.4009	10.6	43.1	69.8	0.3382	7.2	26.0	50.0	-	-	-
	- w/ Corrected-SegMap	0.4616	28.6	63.1	80.8	0.3826	21.6	46.8	60.8	-	-	-
(b2)	SegMap + SIFT + RANSAC-CS [91]	0.4162	9.4	38.0	70.4	0.3503	6.0	21.6	50.4	-	-	-
	- w/ Corrected-SegMap	0.4360	25.9	48.2	71.8	0.3651	14.8	30.8	51.6	-	-	-
(b3)	SegMap + SuperPoint + RANSAC-CS [91]	0.4133	14.1	45.5	80.4	0.3701	8.6	34.5	63.5	-	-	-
	- w/ Corrected-SegMap	0.4335	23.5	60.8	85.5	0.3831	14.8	46.8	74.4	-	-	-
(d2)	SegMap + DRMIME [81]	0.1971	0.0	5.5	6.3	0.1893	0.8	3.6	5.6	-	-	-
	- w/ Corrected-SegMap	0.4269	14.1	52.5	83.1	0.3881	16.1	43.2	73.6	-	-	-

for all other datasets.

(c-d) Optimization based methods.

(c1) REMPE [89] assumes the eyeball as an ellipsoid. It first back-projects the 2D retinal correspondences onto the 3D eyeball, and then finds the optimal registration by minimizing the 3D distances of the correspondences. Particle Swarm Optimization is used to find the optimal relative camera pose, and the eyeball's shape and rotation. We use an executable program from the authors, where we have no knowledge on how to correctly warp the manually labeled keypoints. Therefore, its success rates are not reported.

(d1)-(d2) DRMIME [81] is an iterative-searching based method multi-modal images registration. Edges are detected on the image pyramids created from the input images, and the optimal transformation parameters are determined based on Mutual Information (MI) metric. We set five-layer image pyramids and 20% sampling rate in our experiments. In addition to the original scheme (d1) that use images as inputs, we also replace the edge detection step with our vessel segmentation maps as (d2) which yields better results.

(e) Fully network based method.

(e1) DILR [14] pipeline consists of global- and deformable-alignment networks for medical image registration. The networks are trained with negative Normalized Cross Correlation (NCC) loss on the input images. We only implement the affine-alignment network in our experiments.

(f) Our proposed method.

(f1)-(f3) and (f5) are as described in Section 3.4.5. (f4) is based on our previous distortion correction method [90] with only one global parameter d.

Table 3.1 shows the *Dice* values and success rates on all datasets. Overall, our proposed registration pipeline (f5) which has complete distortion correction achieves the highest *Dice*

values and success rates across all datasets.

In detail, when comparing our network (f2) with the keypoint-based methods (b1) & (b3), our method shows advantages on the majority of the metrics, except that our success rates with $threshold = 10$ on UWc datasets is lower than RANSAC-CS. This demonstrates the benefits of the deep neural networks on the outlier rejection task. Next, if comparing (b2) which uses SIFT feature with (b3) which uses SuperPoint Network, it can be observed that (d3) achieve slightly better *Dice* values but much better success rates, especially on UWc dataset which need to align NA images with the heavily distorted UWF peripheral areas. This shows the advantages of the feature detection and description network in finding more robust features for matching and transformation estimation. Besides, the methods (d1) & (e1) which use retinal images as inputs basically fails on this task, while using vessel segmentation maps as input (d2) can greatly improve the registration performance. This shows that the conventional solutions for multi-modal registration (*i.e.*, applying alignment metrics, *e.g.*, NCC and MI, on the original images) are not suitable for multi-modal retinal image registration, but the anatomical structure from retina (*i.e.*, vessels) can greatly alleviate this problem.

When comparing different transformation models of (f1)-(f3), 2nd-order polynomial transformation (f2) achieves the best *Dice* values. The 3D-involved DLT (f3) produces much more successful alignment with $threshold = 10$ but also has more failure cases *i.e.*, with $threshold = 50$. This shows the importance of adopting 3D prior information in aligning UWF with NA retinal images. Nevertheless, DLT might be inappropriate for this task when the actual eyeball shape deviates from the pure-sphere assumption. Finally, by incorporating the proposed distortion correction method in the testing phase (f4) & (f5), the registration performance is boosted in most aspects compared with (f2), which demonstrates the importance of 3D shape prior in accurately aligning UWF with NA images. Our best method (f5) also surpass the DLT model (f3), indicating that we have proposed a more appropriate adoption of 3D information.

When comparing across the UW and UWc datasets, the 5-parameter correction model

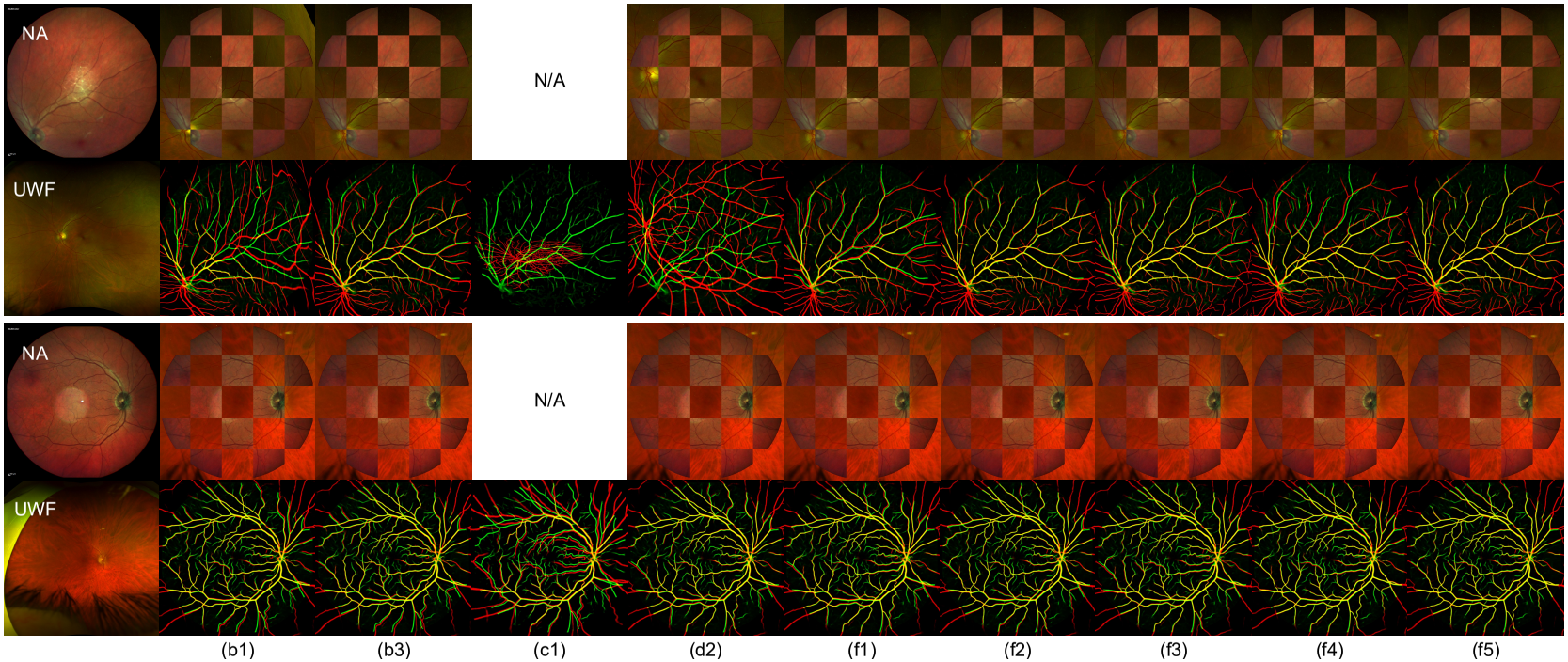


Figure 3.9: Qualitative alignment results. The examples show alignment in peripheral and center areas respectively. In each example, the top row shows the interlaced view of the two aligned images, while the bottom row shows the alignment of their vessels (red and green are from NA and UWF images respectively). (b1) SegMap + SuperPoint + RANSAC (polynomial). (b3) SegMap + SuperPoint + RANSAC-CS [91] (polynomial). (c1) SegMap + SuperPoint + REMPE [89] (3D). (d2) SegMap + DRMIME [81] (Perspective). (f1) Our Network w/o DC (Perspective). (f2) Our Network w/o DC (Polynomial). (f3) Our Network w/o DC (DLT). (f4) Our Network w/ DC (Polynomial, 1 parameter) [90]. (f5) Our Network w/ DC (Polynomial, 5 parameters).

(f5) achieves about 0.014 increase in *Dice* and similar success rates compared to 1-parameter model (f4) on the UW datasets. This indicates that the newly-proposed correction model is able to improve the alignment accuracy for NA image centered on fovea when the alignment is successful, but cannot correct the errors in the failure cases ($threshold > 50$). However, on the UWc datasets which contain more NA images capturing the peripheral retina, the increase in *Dice* advances to about 0.05, and the success rates also have larger improvements. Partial of the failed alignments by (f4) might be caused by the peripheral distortions instead of wrong estimation of inliers, as they can be improved by incorporating more parameters. This shows that our 5-parameter method can find better alignment in peripheral retina with more severe distortions.

Fig. 3.9 shows alignment results on two image pairs from UWc dataset. In the center area alignment, most methods are able to find roughly correct alignments. Nevertheless, the methods using linear transformation (d2, f1) or without distortion correction (b1, b3, f2) have more misalignment. Furthermore, in the example of peripheral area alignment, our method (f5) achieves the best alignment quality. The distortion correction method with only one global parameter (f4) suffers more misalignment in this example, indicating that more parameters are necessary in correcting peripheral distortions.

3.5.4 Ablation Study

Performance Improvement w.r.t. NA Image Position

Fig. 3.10 illustrates the improvements of the 5-parameter correction method (f5) over the 1-parameter method (f4) w.r.t. the NA image's position on retina. Most points show green or cyan colors, indicating improvements on alignment qualities for most cases. Furthermore, when comparing the points around the origin (*i.e.*, NA images centered on fovea) and those peripheral ones (*i.e.*, NA images capturing peripheral retina), the peripheral points show higher saturation than the centered points. This shows that the 5-parameter method is more capable at handling

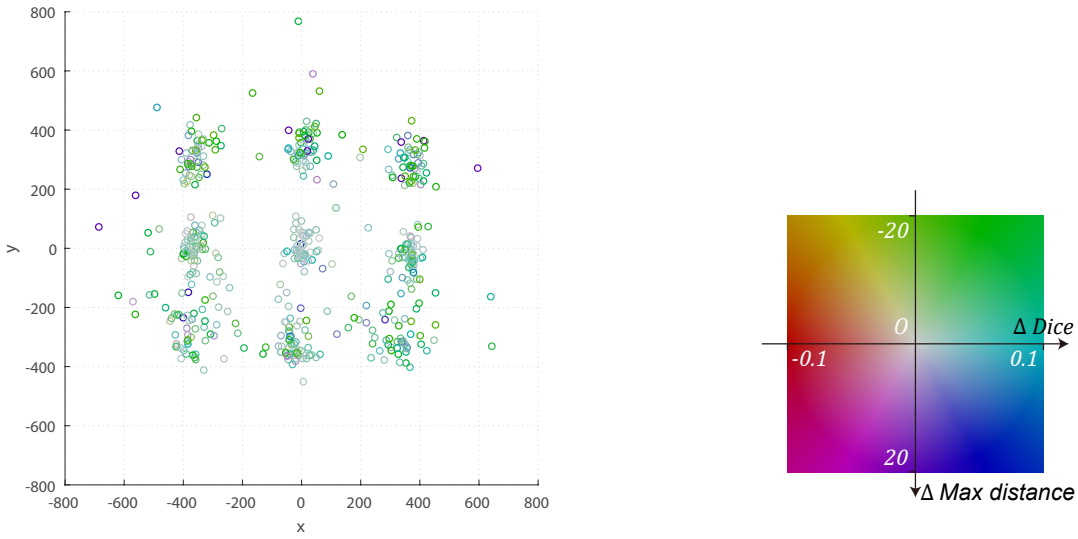


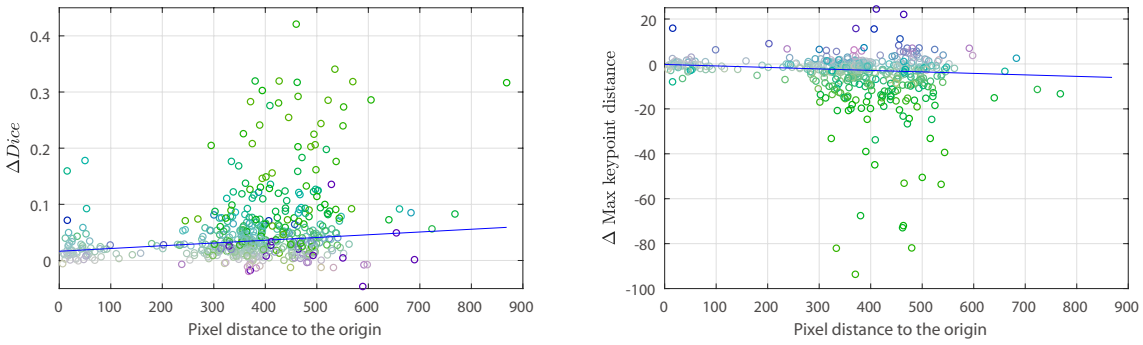
Figure 3.10: Plot of increased values of *Dice* and d_{max} on UWc dataset w.r.t. the NA image's position in the UWF image, when comparing our proposed 5-parameter correction algorithm (f5) over the 1-parameter method (f4) [90]. The origin $(0,0)$ is the UWF's image center and assumed to be fovea. Each circle indicate an image pair. Its location indicates the NA image's center coordinate on retina (in UWF's pixel coordinate) which is derived by warping the NA image center towards UWF modality through affine transformation based on manual labels. Its color legend is shown on the right side. Green color indicates improvements in both *Dice* (increased) and d_{max} (decreased). Purple points generally indicates lower alignment quality, especially due to increased d_{max}.

peripheral images with larger distortions.

In addition, Fig. 3.11 also plots the increased *Dice* and d_{max} of (f5) over (f4) w.r.t. the distance between the fovea and the NA image's center. The majority of image pairs show improvement in *Dice*, while most pairs show similar or lower d_{max}. This can also be verified by the blue lines estimated via robust linear regression, where *Dice* increases and d_{max} decreases as the NA camera looks further away from the fovea.

Network Involvement in Distortion Correction

We compare the performance of distortion correction schemes (f5)-(f7) as summarized in Table 3.2. These schemes involve different number of networks in the searching and correction



(a) (b)

Figure 3.11: Changes of *Dice* and d_{max} w.r.t. the NA image's center distance to the fovea when comparing (f5) over (f4.). The blue lines are estimated on the points by robust linear regression (robustfit function in Matlab) to reduce the influences from outlier points.

process, as described in Section 3.4.5. As observed, with more networks involved, both the success rates and *Dice* values show improvements. In the method using outlier rejection network (f7), the corrected keypoint positions might have helped the network to find more true inliers and exclude false inliers, which therefore achieves better performance than (f6). Furthermore, if feature detection is involved (f5), the network would be able to extract features from two images with similar distortions, which might have further benefited the following steps and improved the performance. Meanwhile, even though the computation time per iteration increases as more networks are involved, the average computation time for each image pair decreases, because much fewer iterations are needed to reach the optimal point.

Performance Improvement with Corrected Input Maps

Table 3.2 also compares the registration performance when using the original or distortion-corrected vessel segmentation maps as input for feature detection and image warping. The corrected maps are based on the optimal 5 parameters estimated by our proposed method. RANSAC and RANSAC-CS (b1)-(b3) as well as DRMIME (d2) are included in this experiment. As observed, when using the corrected input maps, all methods shows obvious improvements

in registration quality. Especially, the RANSAC method (b1) rivals or exceeds RANSAC-CS (b3) in most evaluation metrics after adopting the corrected segmentation maps, which proves the effectiveness of distortion correction for alignment performance improvement.

It should also be noted that, the dramatic boost of DRMIME's performance (d2) is not only caused by the reduction of peripheral distortions, but also resulted from the the knowledge of the NA's image center $\mathbf{X}_{cen}$ from Algorithm 4 where the initial translation between two images are reduced and DRMIME can find the optimal transformation more easily.

3.6 Conclusion

In this paper, we propose a distortion correction method to reduce the peripheral distortions in UWF images for UWF-to-NA retinal image alignment. The correction model functions by remapping the UWF image based on a similar camera pose as the NA image which is described by 5 parameters. Along with the correction function, we set up a registration pipeline consisting of CNN, and incorporate the distortion correction method into the pipeline during testing phase. A two-step searching algorithm first finds the globally optimal distance on the training images between the NA camera and the eyeball. Then, it iteratively optimizes the NA camera pose for each image pairs to achieve the best alignment performance. Experimental results show that the proposed method achieves the best alignment quality, as well as the capability of the distortion correction function in improving UWF-to-NA registration performance.

In the future, we would like to further exploit the alignment between multiple NA and single UWF images, as well as the joint alignment of over three modalities. 3D information will likely play a crucial role in these tasks.

Chapter 3, in full, has been submitted for publication of the material as it may appear in IEEE Transactions on Image Processing, 2022, Junkang Zhang; Yiqian Wang; Fritz Gerald P. Kalaw; Melina Cavichini-Cordeiro; Dirk-Uwe G. Bartsch; William R. Freeman; Truong Q.

Nguyen; Cheolhong An, IEEE, 2022. The dissertation author was the primary investigator and author of this material.

4 3D Eyeball Shape Estimation for Ultra-Widefield and Narrow-Angle Retinal Image Alignment

4.1 Introduction

As shown in Chapter 3, the alignment performance between Ultra-Widefield (UWF) and Narrow-Angle (NA) images achieves large improvements when 3D eyeball shape information is incorporated into the alignment process. The spherical assumption of the eyeball shape helps to reduce the non-linear 2D distortions in UWF images which cannot be solved by simple 2D-to-2D global transformation models. However, this approximation still deviates from the actual eyeball shape which is more complex than a sphere. Therefore, there will still exist misalignments even if we have estimated accurate NA camera poses and very complex 2D-to-2D transformation models. In this chapter, we aim to recover a more accurate eyeball shape from the UWF and NA image pair, in order to achieve the best alignment quality between the two images.

There have been several works on the 3D estimation of eyeball curvatures and shapes. However, most of them are based on simple shape functions for the eyeball which are described by only a few global parameters, *e.g.*, spherical, quadratic, or ellipsodial functions. As mentioned in Chapter 3, Ataer-Cansizoglu *et al.* [88] estimate the eyeball shape from multiple NA images

based on planar, spherical and quadratic assumptions, and optimize the shape parameters and multiple camera poses via bundle adjustment. Hernandez-Matas *et al.* [87, 89] jointly estimate the relative camera pose between two retinal images along with the spherical and ellipsodial parameters of the eyeball model via particle swarm optimization. Besides, Chanwimaluang *et al.* [92] estimate retinal curvature from NA image sequences through Structure From Motion which incorporates constraints on ellipsoid surface and lens distortions.

In this chapter, instead of estimating the parameters of a simple 3D functions to describe the eyeball shape, we reconstruct a dense and more accurate scene for the eyeball. Starting from the spherical shape and the estimated NA camera extrinsic parameters by the distortion correction networks in Chapter 3, we first set up an initial 3D mesh with dense vertices over the UWF image. Then, the coordinates of the vertices and the NA camera parameters are jointly and iteratively optimized, with the objective that the reprojected images from the scene based on UWF/NA cameras can exactly match the original 2D input images. In addition, we propose a searching scheme where the iterative optimization process is combined with the distortion correction results from Chapter 3, such that the 3D-to-2D scene projection and 2D-to-2D global alignment can be concatenated to achieve the best alignment quality.

4.2 Proposed Method

4.2.1 Model Setup

Our objective is to reconstruct a dense 3D scene (eyeball shape) for the UWF image, as well as estimate the NA camera's parameters, such that the reprojected 2D images from the scene based on the cameras can be accurately matched with the original input images.

At the starting point, we set up an initial ideal eyeball model in the world coordinate based on stereographic projection, as shown in Fig. 4.1. The eyeball is a pure sphere of radius 1 centered at $(0, 0, 0)$. The UWF camera is located at $(0, 0, \text{-}1)$ with its viewing direction towards

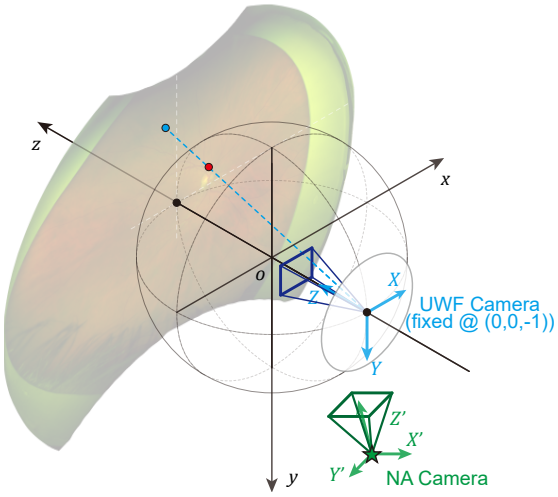


Figure 4.1: 3D eyeball model based on spherical assumption and stereographic projection.

$z =$ -1. We define an initial mesh with vertices $\mathbf{V} = \{\mathbf{v} \in \mathbb{R}^3 | \|\mathbf{v}\|_2 = 1\}$ for UWF imaging area on the eyeball. The vertices $\mathbf{V}$ are arranged in the shape $[W/s+1, W/s+1, 3]$ as shown in Fig. 4.2, where W is the pixel width of the UWF image (assumed to have square shape), and $s \in \mathbb{Z}^+$ is the sampling step of the mesh grid. A vertex $\mathbf{v}$ is located on the pure sphere at

$$\mathbf{v} = \frac{(2fX, 2fY, f^2\text{-}X^2\text{-}Y^2)}{f^2+X^2+Y^2}, \tag{4.1}$$

where $f = 2$ is the focal length of the UWF camera. $[X, Y]$ is the vertex's 2D location in the UWF image obtained by

$$[X, Y, f]^T = [m/\gamma, n/\gamma, f]^T, \tag{4.2}$$

where γ is a known scaling factor between the pixel coordinate and image coordinate, and $[m, n]$ is the corresponding location in the UWF pixel coordinate. Readers can refer to Section 3.3.1 for detailed math derivations.

Then, the UWF mesh reflecting the actual eyeball shape is defined on vertices $\mathbf{V} + \Delta\mathbf{V} = \{\mathbf{v} + \Delta\mathbf{v}\}$, where $\Delta\mathbf{v} \in \mathbb{R}^3$ is the movement of the vertex $\mathbf{v}$ and can have either one or three degrees of freedom which will be discussed later. The NA camera pose is defined by 6 extrinsic

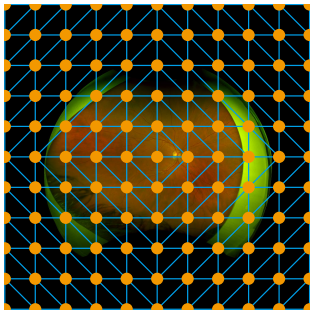


Figure 4.2: An example of a mesh defined on a square UWF image. The image width is $W = 4000$ pixels. The mesh vertices's (orange circles) sampling step is $s = 400$ pixels. The faces (triangles) of the mesh are represented by blue lines.

parameters, *i.e.*, camera position $\mathbf{x}_{cam} \in \mathbb{R}^3$ and orientation $\theta = [\theta_x, \theta_y, \theta_z]$. In addition, we include an intrinsic parameter for the NA camera definition to describe the range of projection (assuming the scene is projected onto a square image), *i.e.*, NA camera field of view α. We can write the reprojected UWF image based on NA camera as

$$I'_{uwf} = \text{project}(I_{uwf}, \mathbf{V} + \Delta\mathbf{V}, \{\mathbf{x}_{cam}, \theta, \alpha\}), \tag{4.3}$$

where project$(\cdot)$ reprojects the 3D scene to the 2D imaging plane defined by the camera, which is implemented in Pytorch3D and able to back-propagate. Finally, in order to achieve our objective, we will jointly optimize both the eyeball shape $\Delta\mathbf{V}$ and the $6+1$ NA camera parameters.

4.2.2 Optimization with Hard Constraint on UWF Reconstruction

An intuition here is that, when the scene is being altered during optimization, we hope that the reprojected UWF image by the UWF camera still remains identical to the original UWF image. To achieve this goal, the movement of a vertex $\Delta\mathbf{v}$ should be restricted on the UWF camera's projection ray of the corresponding pixel. The viewing direction of the ray is

$$\vec{\mathbf{n}}_u = \text{norm}(\mathbf{v} - [0, 0, -1]^T). \tag{4.4}$$

Therefore, the vertex's movement is defined as

$$\Delta \mathbf{v} = \Delta v \cdot \vec{\mathbf{n}}_u, \tag{4.5}$$

where $\Delta v \in \mathbb{R}$ has one degree of freedom, and all movements $\Delta V = \{\Delta v\}$ is an array with shape $[W/s+1, W/s+1]$.

The optimization process for ΔV and NA camera parameters is based on two loss terms, *i.e.*, photometric consistency loss and smoothness loss. On one hand, the reprojected UWF image by the NA camera should match the NA image. To this end, we extract vessel maps from both retinal modalities, reproject the UWF vessel map instead of its color pixels, and measure the alignment quality between the reprojected UWF vessel map with the NA vessel map. We use photometric consistency loss to evaluate the alignment quality which is defined as

$$L_{pc} = \text{MSE}(\text{project}(I_{uwf\text{-}s}, \mathbf{V} + \Delta \mathbf{V}, \{\mathbf{x}_{cam}, \theta, \alpha\}), I_{na\text{-}s}). \tag{4.6}$$

On the other hand, the estimated scene $\mathbf{V} + \Delta \mathbf{V}$ should be smooth without abrupt changes such as sharp edges or corners. Therefore, we apply the smoothness on the estimated scene as

$$L_{sm}(\Delta V) = \underset{i,j}{\text{mean}} \left((\Delta V_{i,j} - \Delta V_{i+1,j})^2 \right) + \underset{i,j}{\text{mean}} \left((\Delta V_{i,j} - \Delta V_{i,j+1})^2 \right), \tag{4.7}$$

where i and j are indices in the array of ΔV. Finally, the total loss is defined as

$$L_{hard} = L_{pc} + \lambda_{sm} L_{sm}. \tag{4.8}$$

4.2.3 Optimization with Soft Constraint on UWF Reconstruction

However, we find that it is difficult for the optimization process with the hard constraint to achieve the best alignment, which is likely due to the limited direction of vertex's movement.

Therefore, we lift the hard constraint on identical UWF image reconstruction, and add a soft constraint for it instead during optimization. Meanwhile, we allow the vertex to be adjusted in all orientations (*i.e.*, three degrees of freedom) instead of in a restricted direction. In other words, we directly optimize the array of $\Delta\mathbf{V} = \{\Delta\mathbf{v}\}$ with shape $[W/s+1, W/s+1, 3]$.

The loss terms for the optimization process consist of three parts. First, we use the same photometric consistency loss from Eq. (4.6) between the reprojected and the original vessel maps. Second, the smoothness loss is defined as

$$L_{sm}(\Delta\mathbf{V}) = \underset{i,j,k}{\text{mean}}\left((\Delta\mathbf{V}_{i,j,k} - \Delta\mathbf{V}_{i+1,j,k})^2\right) + \underset{i,j,k}{\text{mean}}\left((\Delta\mathbf{V}_{i,j,k} - \Delta\mathbf{V}_{i,j+1,k})^2\right), \tag{4.9}$$

where i, j and k are indices in the array of $\Delta\mathbf{V}$. Next, we define a direction deviation loss to penalize the inconsistent moving direction of $\Delta\mathbf{v}$ from the direction of the ideal projection ray $\vec{\mathbf{n}}_u$ in the UWF camera

$$L_{dir} = \underset{\forall\Delta\mathbf{v}}{\text{mean}} \|\Delta\mathbf{v} - \langle\Delta\mathbf{v}, \vec{\mathbf{n}}_u\rangle \cdot \vec{\mathbf{n}}_u\|_2^2 \tag{4.10}$$

which acts as the soft constraint for UWF reconstruction. A smaller L_{dir} value indicates that the reprojected UWF image based on UWF camera is more similar with the original input. Finally, the total loss is defined as

$$L_{hard} = L_{pc} + \lambda_{sm} L_{sm} + \lambda_{dir} L_{dir}. \tag{4.11}$$

4.2.4 Optimization Process

Assume that a set of sparse correspondence $\mathbf{p} = [..., (m_i, n_i, u_i, v_i), ...] \in \mathbb{R}^{N\times 4}$ between the two images as well as their weights $\mathbf{w} \in \mathbb{R}^{N\times 1}$ are available (*e.g.*, those estimated by the SuperPoint and outlier rejection networks in Chapter 3), we first need to find the initial estimate of the $6+1$ NA camera parameters before optimizing $\Delta\mathbf{V}$ which will otherwise move in the wrong direction. A simple method to achieve this is to optimize the NA camera parameters by minimizing the average weighted distance between NA keypoints $[u_i, v_i]$ and reprojected UWF

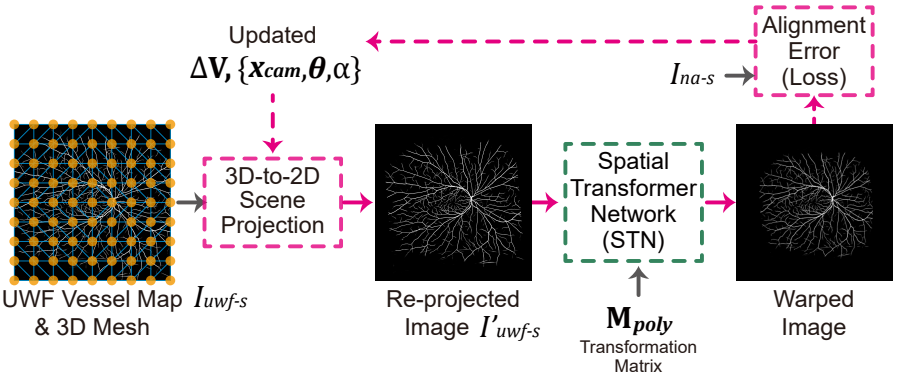


Figure 4.3: Incorporating the proposed scene optimization process into existing methods.

keypoints

$$L_{init\text{-}cam} = \underset{i}{\text{mean}}\, w_i \left\| [u_i, v_i] - \text{project}\big([m_i, n_i], \mathbf{V}, \{\mathbf{x}_{cam}, \theta, \alpha\}\big) \right\|_2^2 \tag{4.12}$$

where the project$(\cdot)$ is similar to Algorithm 1 in Chapter 3. Afterwards, both $\Delta\mathbf{V}$ and NA camera parameters can be jointly and iteratively optimized by minimizing Eq. (4.11) or (4.8).

4.2.5 Incorporation into Distortion Correction Network

A better option to apply our proposed algorithm is to incorporate it into the Distortion Correction process of Chapter 3, since it uses the more complex 2nd-order polynomial transformation and has already achieved good alignment quality. The incorporated optimization process is illustrated in Fig. 4.3. The initial inputs are the estimated transformation matrix $\mathbf{M}_{poly}$ and the 5 correction camera parameters. With the inputs, we can derive the initial estimates of $\mathbf{x}_{cam}$ and $\theta = [\theta_x, \theta_y, 0]$ according to Section 3.3.2, while the camera field of view is initialized as

$$\alpha \leftarrow 2 \cdot \tan^{-1} \frac{W/(2\gamma')}{2 + f'}. \tag{4.13}$$

During iterative optimization, we first output a reprojected UWF vessel map based on the current estimate of the scene and the NA camera parameters. Then the reprojected UWF vessel map is warped by $\mathbf{M}_{poly}$. Finally, we compute the loss using the warped UWF vessel map and the NA vessel map, and update $\Delta\mathbf{V}$ and NA camera parameters.

Table 4.1: Average *Dice* Values (Standard Deviation) on UWc Datasets

Method	UWc1-2 (255)	UWc2-1 (250)
Input (Distortion Correction Network)	0.5674 (0.1062)	0.5176 (0.1298)
Deformable Alignment Network (Section 3.4.1)	0.6125 (0.1209)	0.5707 (0.1454)
Scene Reconstruction (Soft Constraint)	0.6850 (0.1248)	0.6281 (0.1582)

4.3 Experiments

4.3.1 Settings

We use the UWc dataset in our experiment which has been introduced in Chapter 3. We only use the center 2000×2000 area of the UWF image (*i.e.*, $W = 2000$). The rasterization size in the image projection is 1000×1000, and the MC images are expanded to the same size. For each image pair, we set up a mesh with $s = 8$, *i.e.*, 126×126 vertices in $\mathbf{V}$ and $\Delta\mathbf{V}$, and initialize the NA camera parameters by the estimation results of the distortion correction model from Chapter 3. The vessel segmentation map as well as the polynomial transformation matrix are pre-computed and kept unchanged during optimization.

During optimization, we use Adam optimizer [72] to update the ΔV (learning rate $1e$-2) and 6+1 NA camera parameters (learning rate $1e$-4). The weights for each loss terms are $\lambda_{sm} = 1e5$ and $\lambda_{dir} = 1e2$ under soft constraint. Under soft constraint, the parameters are updated by 200 iterations, and the optimization process will terminate earlier if the increased amount of current *Dice* value over the one of 20 iterations ago is smaller than $1e-3$. *Dice* values are computed between the NA and the warped UWF vessel maps to evaluate the alignment quality. The algorithm is implemented in Pytorch and Pytorch3D, and tested on a server with GTX 1080 Ti graphics cards.

4.3.2 Results

Table 4.1 shows the *Dice* values between the NA and the warped UWF vessel maps. We use the deformable alignment network trained in Section 3.4.1 as the comparison method

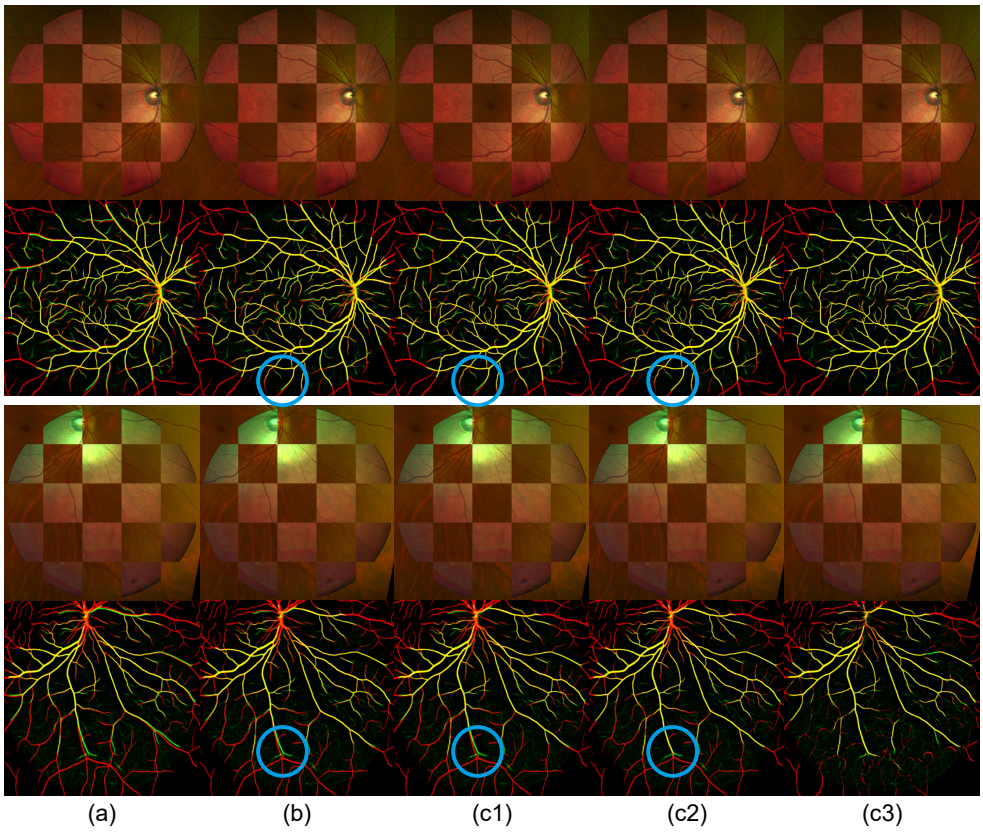
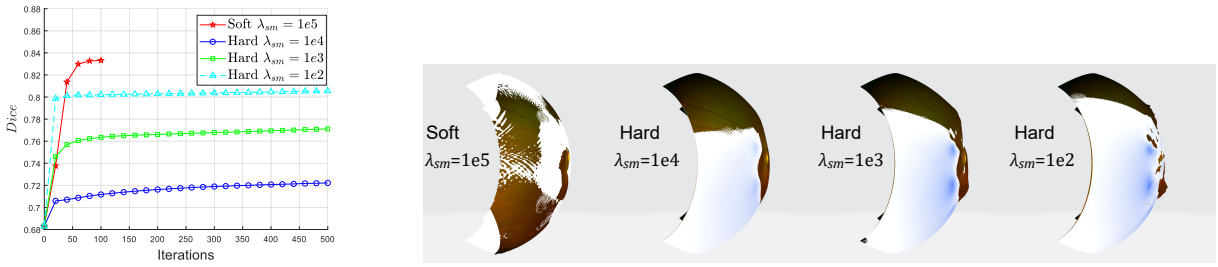


Figure 4.4: Qualitative alignment results on two pair of input images. In each example, the top row shows the interlaced view of the two aligned images, while the bottom row shows the alignment of their vessels (red and green are from UWF and NA images respectively). (a) Input image aligned with the global transformation by the distortion correction network of Chapter 3. (b) Fine alignment by the deformable alignment network. (c1)-(c3) Alignment by our proposed eyeball shape estimation method with soft constraint on UWF reconstruction, where $\lambda_{sm} = 1e6$, $1e5$ and $1e4$ respectively.

which makes one-time predictions of optical flow fields for 2D-to-2D warping. As observed, the proposed eyeball scene reconstruction method using the soft constraint has advantages over the deformable network with over 0.05 increase in *Dice* values.

Figure 4.4 shows qualitative results from different methods, as well as the scene reconstruction method with different settings. The proposed method (c2) shows better alignment quality than the deformable network (b) in details indicated by blue circles. Besides, in (c1)-(c3), we compare the results optimized under different weights λ_{sm} for the smoothness loss term. When



(a) (b)

Figure 4.5: Comparison of soft ($\lambda_{sm} = 1e5$) and hard ($\lambda_{sm} = 1e4$, $1e3$ & $1e2$) constraints for UWF image reconstruction during scene optimization. (a) Plot of *Dice* w.r.t. optimization iterations. (b) Reconstructed eyeball shapes under various optimization settings, from a same view. In the images, the white or blue surfaces represent the pure sphere (*i.e.*, initial estimation of the shape), while the UWF image act as the UV map which is painted on the estimated scenes.

using a smaller weight of $1e4$ in (c3), the unmatched UWF vessels are reduced in width, which indicates unfavorable abrupt local changes in the estimated scene. On the other hand, when using a larger weight of $1e6$ in (c1), the alignment performance becomes less competitive. Finally, the setting of $\lambda_{sm} = 1e5$ achieves the balance between the best alignment quality and smooth shape of the reconstructed scene, which is therefore adopted for our experiment.

Figure 4.5 shows a comparison of using soft and hard constraint when optimizing on one image pair. As can be seen, the scheme using soft constraint is able to reach the highest *Dice* score in fewer iterations while keeping the reconstructed eyeball shape smooth. By contrast, the schemes adopting hard constraint need much more iterations before achieving high *Dice* scores. Moreover, achieving higher *Dice* scores by reducing smoothness weight λ_{sm} leads to more uneven eyeball shapes. Therefore, with regard to improving retinal image alignment quality, the scheme using soft constraint is a better choice.

4.4 Conclusion

In this chapter, in order to achieve higher performance for UWF-to-NA retinal image alignment, we set up an iterative optimization algorithm to estimate finer eyeball shape defined

on a dense 3D mesh, such that the reprojected UWF image can be accurately matched with the NA image. The objective for the optimization process is to generate reprojected UWF vessel structures based on UWF/NA cameras which remain identical to the original UWF/NA images, while the estimated eyeball shape maintains smoothness. We incorporate the constraint on UWF vessel map reconstruction as either a soft or a hard constraint in the algorithm. Besides, the proposed iterative optimization process is concatenated with the global alignment results from the distortion correction networks of Chapter 3, so it becomes similar to a fine alignment process which reduces the misalignment. Through our experiments, we demonstrate the effectiveness of the proposed algorithm, and discuss the usage of the soft and hard constraints.

Chapter 4, in part, is currently being prepared for submission for publication of the material. Junkang Zhang; Yiqian Wang; Fritz Gerald P. Kalaw; Melina Cavichini-Cordeiro; Dirk-Uwe G. Bartsch; William R. Freeman; Truong Q. Nguyen; Cheolhong An. The dissertation author was the primary investigator and author of this material.

5 Conclusion

In this thesis, we present two pipelines to handle the main challenges in multi-modal retinal image registration.

To solve the inconsistency of multi-modal anatomical retinal structures, we proposed to extract the retinal vasculature as a consistent signal between multi-modal retinal images. The vasculature maps can function as the inputs for feature detection and description in the global alignment, as well as the guidance for training the deformable registration network. In addition, to the best of our knowledge, we first proposed the two-step coarse-to-fine registration pipeline completely based on deep neural networks for multi-modal retinal image registration. The coarse alignment network consists of networks of vessel segmentation, feature detection and description, and outlier rejection. In the learning process, the vessel segmentation networks are trained in conjunction with the deformable registration network, without the needs for segmentation ground-truths.

We also proposed a distortion correction module to reduce the misalignment caused by the stereographic projection in UWF, where the correction module remaps the UWF image to a new camera which shares similar extrinsic parameters with the Narrow-Angle (NA) camera. The remapping functions is based on the assumptions of a spherical eyeball shape and a fixed UWF camera pose. The new camera pose is defined by five parameters which are iteratively optimized by minimizing the alignment error. Moreover, the distortion correction module is incorporated into a global registration network to benefit the feature detection and matching process which

will lead to better alignment performance.

Following the distortion correction module, we further remove the restriction of spherical shape assumption, and represent the eyeball shape with a dense 3D mesh which is to be optimized. The objective is to estimate a smooth 3D scene along with the NA camera parameters, such that the reprojected 2D images from the 3D scene based on either UWF/NA camera can be accurately matched with the original input images. In addition, the optimization process is concatenated with the global alignment results from the distortion correction network, so it can work as a fine alignment step to further reduce alignment errors in UWF-to-NA retinal image registration.

In the future, we will further improve and extend our work in the following aspects. First, in recent advancements of general image registration, transformers [93] have been proved effective in image matching in various scenarios, and thus could be introduced into the multi-modal retinal image registration task to help in estimating more accurate correspondence. Second, we will explore the possibilities of finding analytical solutions of the 3D eyeball shape information from dense/sparse correspondence, so that the iterative optimization process can be removed or accelerated. Next, we can align multiple NA images to the UWF image simultaneously which will introduce more constraints and lead to more accurate estimate of the 3D scene. Finally, we will take into consideration other types of distortions introduced in the retinal imaging process, *e.g.*, refractions, eyeball movements.

Bibliography

[1] 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.

[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] E. T. D. R. S. R. Group, "Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified airlie house classification: Etdrs report number 10," Ophthalmology, vol. 98, no. 5, pp. 786–806, 1991.

[4] A. Hassenstein and C. H. Meyer, "Clinical use and research applications of heidelberg retinal angiography and spectral-domain optical coherence tomography–a review," Clinical & experimental ophthalmology, vol. 37, no. 1, pp. 130–143, 2009.

[5] P. A. Keane and S. R. Sadda, "Retinal imaging in the twenty-first century: State of the art and future directions," Ophthalmology, vol. 121, no. 12, pp. 2489–2500, 2014.

[6] A. Nagiel, R. A. Lalane, S. R. Sadda, and S. D. Schwartz, "Ultra-widefield fundus imaging: a review of clinical applications and future trends," Retina, vol. 36, no. 4, pp. 660–678, 2016.

[7] Department of Ophthalmology and Visual Sciences, University of Iowa Health Care, "Fluorescein angiography." [Online]. Available: https://medicine.uiowa.edu/eye/patient-care/imaging-services/fluorescein-angiography

[8] ——, "Indocyanine green angiography." [Online]. Available: https://medicine.uiowa.edu/eye/patient-care/imaging-services/indocyanine-green-angiography

[9] M. Yung, M. A. Klufas, and D. Sarraf, "Clinical applications of fundus autofluorescence in retinal disease," International journal of retina and vitreous, vol. 2, no. 1, pp. 1–25, 2016.

[10] 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.

[11] C. Studholme, D. Hill, and D. Hawkes, "An overlap invariant entropy measure of 3d medical image alignment," Pattern Recognition, vol. 32, no. 1, pp. 71 – 86, 1999.

[12] M. A. Viergever, J. A. Maintz, S. Klein, K. Murphy, M. Staring, and J. P. Pluim, "A survey of medical image registration – under review," Medical Image Analysis, vol. 33, pp. 140 – 144, 2016.

[13] G. Balakrishnan, A. Zhao, M. R. Sabuncu, A. V. Dalca, and J. Guttag, "An unsupervised learning model for deformable medical image registration," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 9252–9260.

[14] 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.

[15] M. C. H. Lee, O. Oktay, A. Schuh, M. Schaap, and B. Glocker, "Image-and-spatial transformer networks for structure-guided image registration," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, 2019, pp. 337–345.

[16] DICOM Standards Committee. (2015) Digital imaging and communications in medicine (dicom) supplement 173: Wide field ophthalmic photography image storage sop classes. [Online]. Available: https://www.dicomstandard.org/News-dir/ftsup/docs/sups/sup173.pdf

[17] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," CoRR, vol. abs/1508.06576, 2015. [Online]. Available: http://arxiv.org/abs/1508.06576

[18] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in Computer Vision – ECCV 2016, 2016, pp. 694–711.

[19] J. J. Yu, A. W. Harley, and K. G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in Computer Vision – ECCV 2016 Workshops, 2016, pp. 3–10.

[20] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 337–33 712.

[21] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016, pp. 467–483.

[22] M. Felsberg and G. Sommer, "The monogenic signal," IEEE Transactions on Signal Processing, vol. 49, no. 12, pp. 3136–3144, 2001.

[23] 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," Biomed. Opt. Express, vol. 9, no. 2, pp. 410–422, 2018.

[24] Álvaro S. Hervella, J. Rouco, J. Novo, and M. Ortega, "Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement," Procedia Computer Science, vol. 126, pp. 97 – 104, 2018.

[25] 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.

[26] 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.

[27] 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 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1046–1053.

[28] 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.

[29] 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), 2018, pp. 2857–2862.

[30] M. Hernandez, G. Medioni, Z. Hu, and S. Sadda, "Multimodal registration of multiple retinal images based on line structures," in 2015 IEEE Winter Conference on Applications of Computer Vision, 2015, pp. 907–914.

[31] D. Motta, W. Casaca, and A. Paiva, "Vessel optimal transport for automated alignment of retinal fundus images," IEEE Transactions on Image Processing, vol. 28, no. 12, pp. 6154–6168, 2019.

[32] T. Chanwimaluang, Guoliang Fan, and S. R. Fransen, "Hybrid retinal image registration," IEEE Transactions on Information Technology in Biomedicine, vol. 10, no. 1, pp. 129–142, 2006.

[33] 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), 2018, pp. 1449–1453.

[34] J. Lee, P. Liu, J. Cheng, and H. Fu, "A deep step pattern representation for multimodal retinal image registration," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5076–5085.

[35] M. Arikan, A. Sadeghipour, B. Gerendas, R. Told, and U. Schmidt-Erfurt, "Deep learning based multi-modal registration for retinal imaging," in Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, 2019, pp. 75–82.

[36] L. Ding, A. E. Kuriyan, R. S. Ramchandran, C. C. Wykoff, and G. Sharma, "Weakly-supervised vessel detection in ultra-widefield fundus photography via iterative multi-modal registration and learning," IEEE Transactions on Medical Imaging, pp. 1–1, 2020.

[37] G. Luo, X. Chen, F. Shi, Y. Peng, D. Xiang, Q. Chen, X. Xu, W. Zhu, and Y. Fan, "Multimodal affine registration for icga and mcsl fundus images of high myopia," Biomed. Opt. Express, vol. 11, no. 8, pp. 4443–4457, 2020.

[38] Y. Tian, Y. Hu, Y. Ma, H. Hao, L. Mou, J. Yang, Y. Zhao, and J. Liu, "Multi-scale u-net with edge guidance for multimodal retinal image deformable registration," in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), 2020, pp. 1360–1363.

[39] J. Zhang, C. An, J. Dai, M. Amador, D. 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), 2019, pp. 839–843.

[40] 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, "A segmentation based robust deep learning framework for multimodal retinal image registration," in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 1369–1373.

[41] C. Harris and M. Stephens, "A combined corner and edge detector," in Proc. of Fourth Alvey Vision Conference, 1988, pp. 147–151.

[42] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vision, vol. 60, no. 2, p. 91–110, 2004.

[43] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," Commun. ACM, vol. 24, no. 6, p. 381–395, 1981.

[44] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 118–126.

[45] D. P. Vassileios Balntas, Edgar Riba and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in Proceedings of the British Machine Vision Conference (BMVC). BMVA Press, September 2016, pp. 119.1–119.11.

[46] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac — differentiable ransac for camera localization," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2492–2500.

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

[48] Xufeng Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3279–3286.

[49] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, "Lf-net: Learning local features from images," in Advances in Neural Information Processing Systems 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 6234–6244.

[50] X. Shen, C. Wang, X. Li, Z. Yu, J. Li, C. Wen, M. Cheng, and Z. He, "Rf-net: An end-to-end image matching network based on receptive field," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8124–8132.

[51] I. Rocco, R. Arandjelović, and J. Sivic, "Convolutional neural network architecture for geometric matching," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 11, pp. 2553–2567, 2019.

[52] I. Rocco, R. Arandjelovic, and J. Sivic, "End-to-end weakly-supervised semantic alignment," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6917–6925.

[53] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in Computer Vision - ECCV 2004, T. Pajdla and J. Matas, Eds., 2004, pp. 25–36.

[54] 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1647–1655.

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

[56] J. Hur and S. Roth, "Iterative residual refinement for joint optical flow and occlusion estimation," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5747–5756.

[57] S. Meister, J. Hur, and S. Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," in AAAI Conference on Artificial Intelligence, 2018, pp. 7251–7259.

[58] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in Advances in Neural Information Processing Systems 28, 2015, pp. 2017–2025.

[59] 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, 2017, pp. 204–212.

[60] 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, vol. 38, no. 8, pp. 1788–1800, 2019.

[61] S. Zhao, Y. Dong, E. Chang, and Y. Xu, "Recursive cascaded networks for unsupervised medical image registration," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10 599–10 609.

[62] S. Zhao, T. Lau, J. Luo, E. I.-C. Chang, and Y. Xu, "Unsupervised 3d end-to-end medical image registration with volume tweening network," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 5, pp. 1394–1404, 2020.

[63] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," IEEE Transactions on Medical Imaging, vol. 22, no. 8, pp. 986–1004, 2003.

[64] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251.

[65] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 2015, pp. 234–241.

[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2014.

[67] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust vessel segmentation in fundus images," International journal of biomedical imaging, vol. 2013, 2013.

[68] J. Staal, M. D. Abramoff, 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.

[69] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Deep retinal image understanding," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, 2016, pp. 140–148.

[70] C. P. Bridge, "Introduction to the monogenic signal," CoRR, vol. abs/1703.09199, 2017.

[71] S. Hajeb Mohammad Alipour, H. Rabbani, and M. R. Akhlaghi, "Diabetic retinopathy grading by digital curvelet transform," Computational and mathematical methods in medicine, vol. 2012, pp. 1607–1614, 2012.

[72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," CoRR, vol. abs/1412.6980, 2014.

[73] G. Azzopardi, N. Strisciuglio, M. Vento, and N. Petkov, "Trainable cosfire filters for vessel delineation with application to retinal images," Medical Image Analysis, vol. 19, no. 1, pp. 46 – 57, 2015.

[74] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, S. 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.

[75] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in Medical Image Computing and Computer-Assisted Intervention — MICCAI'98, W. M. Wells, A. Colchester, and S. Delp, Eds., 1998, pp. 130–137.

[76] K. Zuiderveld, Contrast Limited Adaptive Histogram Equalization. USA: Academic Press Professional, Inc., 1994, p. 474–485.

[77] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, 2005, pp. 886–893 vol. 1.

[78] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," Medical Image Analysis, vol. 57, pp. 226–236, 2019.

[79] T. C. Mok and A. C. Chung, "Fast symmetric diffeomorphic image registration with convolutional neural networks," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4643–4652.

[80] J. Wang and M. Zhang, "Deepflash: An efficient network for learning-based medical image registration," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4443–4451.

[81] A. Nan, M. Tennant, U. Rubin, and N. Ray, "Drmime: Differentiable mutual information and matrix exponential for multi-resolution image registration," in Proceedings of the Third Conference on Medical Imaging with Deep Learning, vol. 121, 2020, pp. 527–543.

[82] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in Computer Vision – ECCV 2018, 2018, pp. 179–196.

[83] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in Proceedings of the 35th International Conference on Machine Learning, vol. 80, 2018, pp. 531–540.

[84] C. Qin, B. Shi, R. Liao, T. Mansi, D. Rueckert, and A. Kamen, "Unsupervised deformable registration for multi-modal images via disentangled representations," in Information Processing in Medical Imaging, 2019, pp. 249–261.

[85] 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," IEEE Transactions on Image Processing, vol. 30, pp. 3167–3178, 2021.

[86] 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, 2022.

[87] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, and A. A. Argyros, "Retinal image registration under the assumption of a spherical eye," Computerized Medical Imaging and Graphics, vol. 55, pp. 95–105, 2017.

[88] E. Ataer-Cansizoglu, Y. Taguchi, J. Kalpathy-Cramer, M. F. Chiang, and D. Erdogmus, "Analysis of shape assumptions in 3d reconstruction of retina from multiple fundus images," in 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), 2015, pp. 1502–1505.

[89] C. Hernandez-Matas, X. Zabulis, and A. A. Argyros, "Rempe: Registration of retinal images through eye modelling and pose estimation," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 12, pp. 3362–3373, 2020.

[90] 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.

[91] L. Ding, T. D. Kang, A. E. Kuriyan, R. S. Ramchandran, C. C. Wykoff, and G. Sharma, "Combining feature correspondence with parametric chamfer alignment: Hybrid two-stage registration for ultra-widefield retinal images," IEEE Transactions on Biomedical Engineering, pp. 1–10, 2022.

[92] T. Chanwimaluang and G. Fan, "Constrained optimization for retinal curvature estimation using an affine camera," in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[93] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8918–8927.