NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME

"**T**RANSLATIONAL **E**NGINEERING IN HE**A**LTH AND **M**EDICINE"

Artificial Intelligence in Ophthalmologic Imaging: Joint Learning of Repeatable and Reliable Detectors and Descriptors for Inter-Device Optical Coherence Tomography Image Registration

---

Postgraduate Diploma Thesis

Athanasios E. Zisimopoulos, MD

Supervisor

Dr. Konstantina S. Nikita

Professor in School of Electrical and Computer Engineering
National Technical University of Athens

Co- Supervisor

Dr. Kaveri Thakoor

Assistant Professor of Ophthalmic Science, Department of Ophthalmology
Columbia University Irving Medical Center

Athens, July 2025

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME

"**T**RANSLATIONAL **E**NGINEERING IN HE**A**LTH AND **M**EDICINE"

Artificial Intelligence in Ophthalmologic Imaging: Joint Learning of Repeatable and Reliable Detectors and Descriptors for Inter-Device Optical Coherence Tomography Image Registration

---

Postgraduate Diploma Thesis

Athanasios E. Zisimopoulos, MD

Supervisor

Dr. Konstantina S. Nikita

Professor in School of Electrical and Computer Engineering
National Technical University of Athens

Co- Supervisor

Dr. Kaveri Thakoor

Assistant Professor of Ophthalmic Science, Department of Ophthalmology
Columbia University Irving Medical Center

The postgraduate diploma thesis has been approved by the examination committee on the 4th of December 2024

1st member	2nd member	3rd member
Dr. Konstantina S. Nikita	**Dr. Georgios Stamou**	**Dr. Athanasios Voulodimos**
(Prof. in NTUA)	(Prof. in NTUA)	(Ass. Prof. in NTUA)

Athens, July 2025

Athanasios E. Zisimopoulos, MD

Graduate of the Interdisciplinary Postgraduate Programme,
“Translational Engineering in Health and Medicine”,
Master of Science,
School of Electrical and Computer Engineering,
National Technical University of Athens

Abstract

Optical Coherence Tomography (OCT) is a cornerstone in ophthalmologic imaging, offering visualization of important eye anatomy and notably the retina. Discrepancies between devices due to method of acquisition or differences in resolution and noise pose render automated registration highly challenging. This thesis explores the feasibility and effectiveness of a deep-learning based approach by utilizing the existing framework of the Repeatable and Reliable Detector and Descriptor (R2D2).

Images of the same retina were captured using two distinct modalities. The first was a high end but expensive and inaccessible device that produced clear resolution images of the retinal layers and the second was a portable and affordable modality but produced images that provided less spatial information. Three different datasets were utilized to produce three models that jointly learned repeatable and reliable detectors and descriptors. The first consisted of the existing images after application of random transformations and pairing of the original with the newly derived augmented images. The second utilized roughly aligned images between two different modalities by expert annotation and considered them as equal. This resulted in the creation of paired images of different modalities. The third dataset was a combination of the first two. The models gave the output of dense descriptors for every pixel, repeatability and reliability heatmaps, both of which were used to extract keypoints for registration.

A quantitative and qualitative evaluation of the keypoints derived by the training of the preexisting model on the original was performed. The three models derived by the corresponding dataset *Crafted (C), Threepoint (3P)* and *Omni (O)* demonstrated strengths and disadvantages in different aspects. 3P performed the best quantitatively while C showed the best repeatability maps in the high-quality OCT dataset and O managed to capture keypoints in the portable OCT dataset in a repeatable and reliable manner. However, each model on its own was not able to produce a satisfying registration result based on the traditional approach of Euclidean distance based mutual descriptor matching. A novel fusion model with a keypoint matching approach demonstrated the best results in multimodal image registration.

This thesis provides a demonstration of the ability of unsupervised or semi-supervised keypoint based deep learning framework for inter-device OCT image registration. While current results are promising, challenges remain for the pipeline to be applicable in the clinical setting. Future work in novel matching strategies, automated masking techniques or other image preprocessing steps is required to bridge the gap between deep learning research and translational applications in ophthalmologic imaging.

Keywords: image registration, keypoints, R2D2, multimodal Image registration, OCT, Retina, Ophthalmologic Imaging

Acknowledgments

This thesis would not be possible without the guidance and support of my supervisor, Dr. Konstantina Nikita, professor in the School of Electrical and Computer Engineering at the National and Technical University of Athens. Any expression of thanks would be an understatement for her endeavors to make sure her students are exposed to the most meaningful stimuli within their field of interest.

I would also like to express my sincere gratitude to Dr. Elisa Konofagou, Professor of Biomedical Engineering and Radiology at Columbia University in the city of New York for supervising the completion of my internship and assuring its smooth and uninterrupted progression.

My deepest thanks go to Dr. Kaveri Thakoor, Assistant Professor of Ophthalmic Science in the Department of Ophthalmology at the Columbia University Irving Medical Center and Principal Investigator of the Artificial Intelligence for Vision Science. By welcoming me in her laboratory and sharing with me the remarkable projects her team was conducting, she profoundly shaped my academic interests and provided me with an inspiring unforgettable experience. Of course, I am equally grateful to Ph.D. student Ye Tian, for giving me the opportunity to collaborate in one of his projects and his invaluable guidance.

Special acknowledgements are due to Spyros Atzamoglou, MD, Christina Kappou, MD, Petroula Mitri, MD and Nefeli Paizi, MD who contributed significantly in the annotation process. My deepest thanks also go out to Dr. Konstantinos Mitsis, for his steady guidance and support throughout. My internship experience would not be the same without the most welcoming Student Doctor Angela McCarthy.

Finally, I would like express my gratitude to the entire team of the Master’s programme Translational Engineering in Health and Medicine. Going from night shifts to late night coding sessions has been a surreal experience to which they have contributed significantly. Their role in helping me achieve the goal of bridging the gap between clinical practice and research has been invaluable.

List of Abbreviations

ABBREVIATION	FULL NAME
2D	Two-Dimensional
3D	Three-Dimensional
3P	Threepoint
AMD	Age-related Macular Degeneration
AP	Average Precision
BRB	Blood Retina Barrier
BRIEF	Binary Robust Independent Elementary Features
C	Crafted
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CT	Computed Tomography
D	Descriptor
FAST	Features from Accelerated Segment Test
FIRE	Fundus Image Registration Dataset
GCL	Ganglion Cell Layer
HD-OCT	High-Definition Optical Coherence Tomography
ILM	Internal Limiting Membrane
INL	Inner Nuclear Layer
IPL	Inner Plexiform Layer
LIFT	Learned Invariant Feature Transform
MRI	Magnetic Resonance Imaging
NFL	Nerve Fiber Layer
NMS	Non-Maximum Suppression
OCT	Optical Coherence Tomography
OPL	Outer Plexiform Layer
ONL	Outer Nuclear Layer
ORB	Oriented FAST and Rotated BRIEF Features
O	Omni
pOCT	Portable OCT
R2D2	Repeatable and Reliable Detector and Descriptor
RANSAC	Random Sample Consensus
RPE	Retinal Pigment Epithelium
SD-OCT	Spectral-Domain Optical Coherence Tomography
SfM	Structure-from-Motion
SIFT	Scale Invariant Feature Transform Keypoint
SSD	Sum-of-Squared-Differences
SS-OCT	Swept-Source Optical Coherence Tomography
TD-OCT	Time-Domain Optical Coherence Tomography
UCN	Universal Correspondence Network
VA	Visual Acuity

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Retina, Optical Coherence Tomography, Image Registration

The eyes are the complex sensory organs responsible for vision, capable of capturing light and interpreting it to images. Their intricate anatomy sees the light pass through the tear film, cornea, aqueous humor, lens and vitreous to ultimately reach the retina. This innermost layer of the eye is responsible for translating incoming light to neural signals that the brain deciphers into vision.

In an optimally functioning system, this chain of events enables activities central to human experience of life. Facial recognition, spatial navigation, appreciation of beauty and memory formation are only some that are facilitated. However, those can be forfeited by any pathology of the visual system. Specifically, diseases targeting the retina – Age-related Macular Degeneration (AMD), diabetic retinopathy, macular holes, retinitis pigmentosa and Stargardt's to name a few- hinder the translational ability of the eye and can often lead to irreversible vision loss. Collectively, these retinal conditions account for a significant percentage of world blindness and are projected to rise in prevalence in coming years. [1, 2]

Inarguably, one of the most vital instruments in the clinician's toolbox to remedy the situation is advances in imaging modalities of the retina, with the most important being Optical Coherence Tomography (OCT). OCT conducts a painless and almost instantaneous examination that yields information of the retinal anatomy and in turn of its physiology. This breakthrough has significantly aided in the effort to combat retinal diseases by allowing for earlier intervention through timely diagnosis and an individualized approach based on the imaging.

Despite these advancements, access to high-quality OCT imaging remains limited. It is not uncommon for tertiary hospitals to lack on-sight access to those modalities, more commonly in underserved populations, due to high cost and low portability. The obvious effect is restrictions in timely diagnosis and inability to supply proper medical advice in a timely manner. [3, 4]

To provide an answer to this issue the scientific community is making efforts towards portable and cost-effective OCT devices. The notable efforts have thus far yielded results that while promising, sacrifice image quality to facilitate accessibility. It is thus warranted to explore whether options other than solely further improving hardware are available. One of those can be

the integration of machine learning approaches to enhance the image quality of low cost OCT devices. [5]

Recent advancements in deep learning tasks have allowed for the completion of super-resolution tasks with a focus on medical imaging. Among proposed architectures tailored for said tasks are Convolutional Neural Networks (CNNs) and more recently encoder-decoder models like the U-net that have proven effective in denoising, enhancement and super-resolution pipelines.

However, a limiting step in the development of such models lies in the optimization of the input data. An upstream task in this workflow, to enhance potential results, is image registration. Image registration is the task of spatially aligning two images taken at different times, from different viewpoints or using different modalities to capture them. [6-8]

1.2 Objective of the Thesis

This thesis aims to explore the use of artificial intelligence and deep learning procedures to address the challenging task of retinal OCT image registration across two different modalities. In pursuit of this goal, the Repeatable and Reliable Detector and Descriptor (R2D2) framework was selected as having the potential to identify consistent and discriminative descriptors across image modalities that can lead the registration process. Three different version of the R2D2 model were developed and fine-tuned using different strategies, including a novel dataset incorporating manually aligned image pairs, to investigate which iteration best supports the necessary keypoint extraction to guide multimodal image registration.

The ultimate objective is to evaluate whether these models can produce keypoints capable of enabling a registration pipeline and, if so, which approach offers the best results. By comparing the output between those models this thesis aims to lay the foundation for reliable, repeatable and interpretable registration methods in ophthalmic imaging. The broader aim is to determine whether these models can produce outputs that are reliable and repeatable enough to guide a registration pipeline, even in the absence of explicit transformation information. While the effectiveness of the full registration process remains an open question, the thesis evaluates model performance both quantitatively and qualitatively through keypoint analysis and lays the groundwork for future integration into more comprehensive image alignment and enhancement pipelines.

1.3 Structure of the Thesis

The thesis is organized into 6 chapters. The introductory chapters aim to provide the foundation and basic knowledge regarding the retina, optical coherence tomography, image registration and the basic deep learning framework used in this thesis, the Repeatable and Reliable Detectors and Descriptors. The next chapters outline the processes that took place and their results. Finally, a chapter dedicated to conclusions and possible future improvements. Specifically:

- Chapter 1 is a brief introduction that describes the motivation and rationale behind the thesis and also outlines the overall structure

1.3 Structure of the Thesis

- Chapter 2 provides a basic background along with the necessary literature review regarding the human retina, OCT, Image Registration and Repeatable and Reliable Detectors and Descriptors
- Chapter 3 describes the methods followed, the datasets created and the initial training of the existing model to provide keypoints and descriptors that would facilitate registration
- Chapter 4 is dedicated to the keypoint extraction process and their quantitative and qualitative evaluation
- Chapter 5 reports the final findings after the application of the derived keypoints and descriptors in both the traditional and a model fusion approach
- Chapter 6 summarizes the conclusions drawn and offers potential future continuations for the intermodal OCT image registration problem.

Chapter 2

Background & Literature Review

2.1 Retina Anatomy and Function

The human eye is a highly specialized, fluid filled sphere-like organ, which is responsible for vision. It consists of three primary layers:

- Outer: Consists of the sclera and cornea. They provide clarity for the photons to enter the eye and structural integrity.
- Middle (Uvea): Consists of the iris, ciliary body and choroid, structures essential in vasculature and accommodation.
- Inner: Contains the retina, the neural tissue essential for vision.

The visual process commences as soon as light touches the tear layer that covers the cornea. After it passes through the cornea it travels through the aqueous humor, the lens and the hyaloid. The refractive power of all the aforementioned media, most important being the cornea and lens, guide the light towards a specialized region of the retina, which will be analyzed further later on. The retina is approximately 0.5mm in thickness and lies in the posterior part of the eye. Upon arriving at the retina, photons encounter the photoreceptor cells. These cells have the ability to convert light into electrical signals via a process called phototransduction, a biochemical cascade. This is the step in the visual process that translates the image viewed into a message the human brain can comprehend. It is for this reason that diseases of the retina can have a debilitating effect on vision. [9]

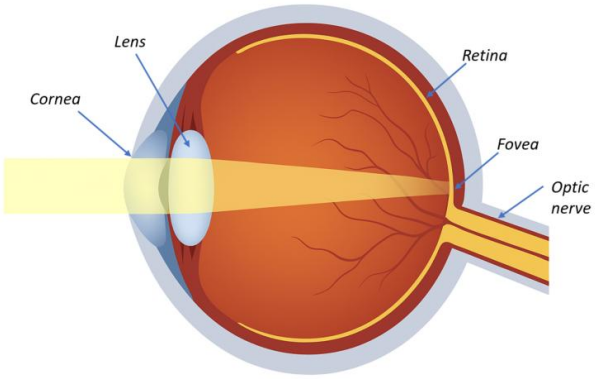

Figure 1 Basic anatomy of human eye

Area centralis, or central retina, is a specific portion of the retina located between the superior and posterior retinal arteries and veins. There lie the macula and fovea, areas of the retina that are responsible for central vision and color perception. They are also areas where retinal

anatomy is significantly different compared to other areas of the retina and where slight changes, due to degenerative diseases or other causes, can cause the greatest effect in visual acuity (VA).

Specifically, the retina can be divided in 9 layers and each one serves a specific purpose. These are the following:

- *Internal Limiting Membrane (ILM):* It is the innermost layer of the retina. It borders the retina from the vitreous humor thus forming a barrier between the two
- *Nerve Fiber Layer (NFL):* This layer contains the axons of ganglion cells which will eventually form the Optic Nerve
- *Ganglion Cell Layer (GCL):* This layer contains the bodies of ganglion cells. These cells receive visual information from the photoreceptor cells.
- *Inner Plexiform Layer (IPL):* This is where synapses (connections) between bipolar and ganglion cells occur
- *Inner Nuclear Layer (INL):* This is a layer mostly occupied by the bodies of bipolar cells. These cells are responsible for transmitting visual information from photoreceptors to ganglion cells
- *Outer Plexiform Layer (OPL):* This is where synapses between bipolar and photoreceptor cells occur
- *Outer Nuclear Layer (ONL):* This is where the bodies of the photoreceptor cell lie. Immediately on its outer part we can also discern the myoid and ellipsoid zone, which signify specific parts of the photoreceptor layer.
- *Retinal Pigment Epithelium (RPE):* This is a layer that supports the retina. Forms the outer Blood Retina Barrier (BRB) and is responsible for metabolic activities of the outer retina.

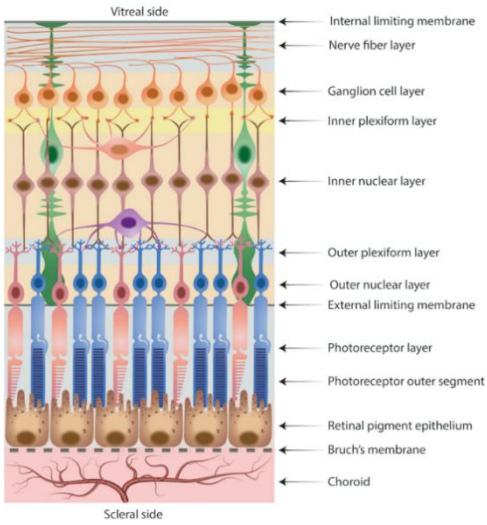

Figure 2 Visual representation of retinal layers

These layers maintain this structure in all areas of the retina, in varying thickness, apart from the fovea centralis. As mentioned before, this is an area 0.35 mm in diameter and is the area responsible for central vision and achieves the highest visual acuity. Here the inner retinal layers are displaced concentrically so that light can reach the photoreceptors with the least possible scattering. Also, this change in anatomy allows for the most efficient packing of cones, cells which specialize in high visual acuity especially in light conditions. [10, 11]

This meticulous arrangement of the neurosensory retina, in combination with the clarity of the optical media is what allows unimpeded flow of photons and consequently the translation of the electric signals into images. However, slight changes in the macular area can significantly impact visual function. These changes more commonly occur in a condition called Age-Related Macular Degeneration (AMD). Most often, this condition can cause scotomas, which are losses of parts of the visual field, drops in central VA or metamorphopsias.

AMD is the leading cause of legal blindness in the industrialized world. It is estimated that by 2040 around 288 million people will be affected by AMD worldwide. It is characterized by accumulation of extracellular deposits, otherwise known as drusen, along with progressive degeneration of photoreceptors and adjacent tissues. These subtle pathological changes need to be closely monitored to ensure the optimal visual outcome. Other than the classic clinical slit lamp examination, the most important imaging modality which can be used to diagnose and monitor AMD disease and progression is Optical Coherence Tomography (OCT). [12]

2.2 Basics of Optical Coherence Tomography

OCT is a completely non-invasive and non-contact imaging technique, especially useful for ophthalmologists, since the transparent optical media of the eye allow for an unobstructed view of key anatomical structures. OCT generates cross-sectional tissue images of high-resolution. Essentially, it mimics tissue biopsy without the necessity of an invasive intervention. Its fast-scanning rates and ability to visualize the image in real time, combined with the higher resolution of OCT compared to other medical imaging modalities, such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI), render OCT a cornerstone for ophthalmic imaging. OCT resolution varies from 20 to 5 μm. [13, 14]

OCT operates based on the principle of emitting light waves at the target tissue and analyzing the delay of the back reflected waves to determine the depth at which the reflection occurred. It uses light near the infrared spectrum and since reflected waves cannot be measured directly, a reference measure is used. This is achieved by splitting the beam into two separate paths: a reference beam toward a reflective surface with known specifications of distance length and a tissue beam directed towards the tissue target as seen in the figure.

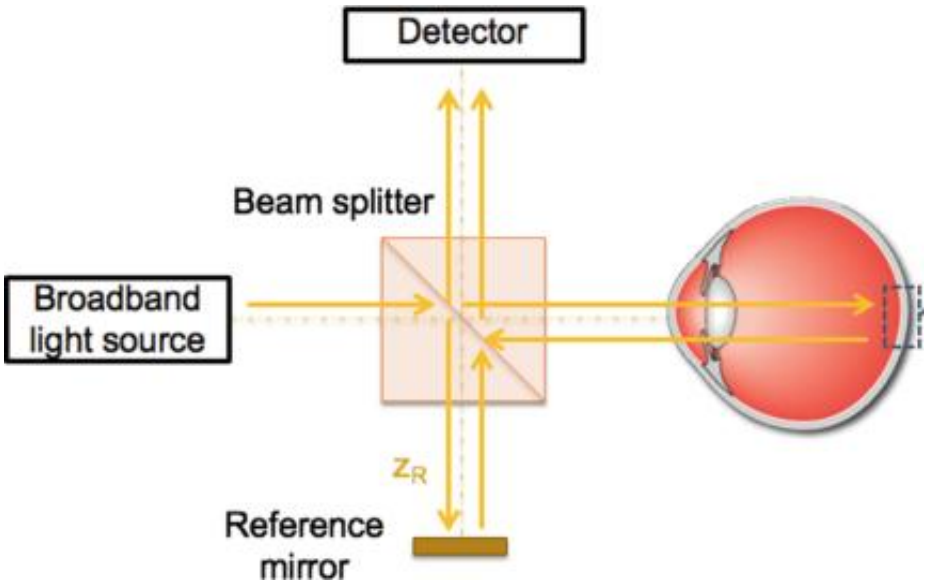

Figure 3
Schematic illustration of the basic principles of OCT. A light source emits light that is split in two beams: one directed to the retina and the other to the reference mirror. The reflection of those two beams is recombined and received by the detector. Interference occurs only when the light path length of the two reflected beams is similar. (6)

The reflected beams from both pathways combine, interfere and reach a detector, which can compare the delay of the back reflection between them. This is achieved by interferometry, which translates to the analysis of how these two beams combine and interfere and in turn provides precise information about reflectivity of the light waves at a specific depth.

An important aspect that needs to be evaluated to achieve a reliable and reproducible reading is coherence. Coherence is a measure of how close in phase two signals are to each other. When it comes to OCT, coherence is significant because interference effects are only detectible when the difference in the path length that light travels between the reference and tissue target beam are within the coherence length of the light source. If these differences are greater than the given coherence length, they cannot be detected by the interferometer. Therefore, it can be easily inferred that low-coherence interferometry is more sensitive in imaging reflections that can differentiate extremely thin layers, such are those of the retina. [15]

Time-Domain OCT (TD-OCT)

Time-domain OCT (TD-OCT) refers to the earliest versions of OCT imaging devices. It is named after the fact that the position of the reference mirror is altered in a manner which produces interference patterns as a function of time. This motion enables the system to scan through the depth of the tissue by almost aligning the optical path lengths of the sample and reference arms at each point in time. Interference is recorded only when the light from the sample and reference arms has traveled nearly identical distances, allowing the detection of backscattered light from specific tissue depths. Because it scans each depth point sequentially, TD-OCT is relatively slow and less efficient than more modern approaches.

Spectral-Domain OCT (SD-OCT)

More advanced OCT devices employ the usage of spectral-domain OCT (SD-OCT). In this instance, the reference mirror's position remains fixed. All light reflections arising from different tissue depths are detected in each lateral scan, significantly reducing the examination time compared to TD-OCT. This is achieved by incorporating a spectrometer, which can capture interference patterns across a broad spectrum of wavelengths all at once. The resulting data is then processed using a Fourier transform so that the received signal can be simplified into the frequencies it is consisted of and thus can plot an accurate representation of the tissues that induced the back scatter. This processing negates the need of a moving part, increases speed and resolution and significantly reduces motion artifacts compared to TD-OCT. These advantages have rendered the SD-OCT the clinical standard for OCT imaging. [15, 16]

Swept-Source OCT (SS-OCT)

Swept-Source OCT (SS-OCT) can be conceptualized as an advancement in SD-OCT. The distinction lies in its use of a tunable narrowband laser that rapidly sweeps across a broad range of wavelengths, typically set around 1050nm. This wavelength achieves greater tissue penetration and can image structures deeper than the RPE, such as the choroid. The spectrometer is replaced with a high speed phtotodetector that records the interference signal as the laser sweeps through the frequencies. Similarly, through Fourier transformation, a depth-resolved reflectivity of the tissues is generated. [17, 18]

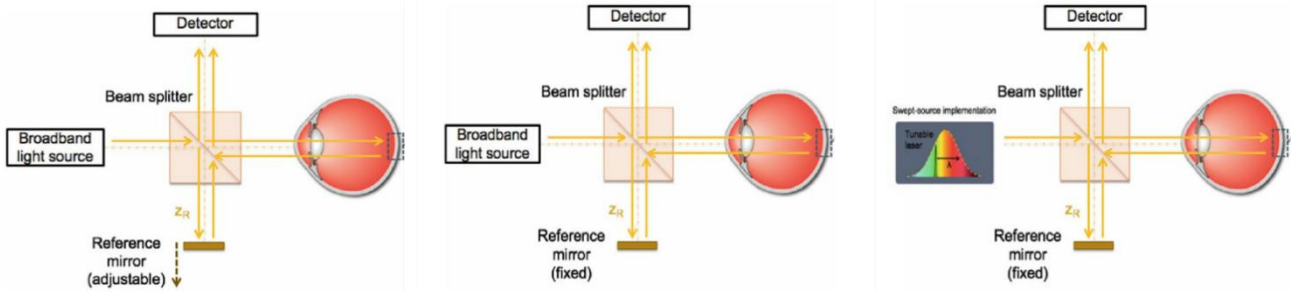

Figure 4
Schematic representation of the basic principles behind Time-Domain, Spectral-Domain and Swept-Source Optical Coherence Tomography (left to right)

These differences, which can be found summarized in the table below, have significantly influenced the conceptual foundation that led to the endeavor that is this thesis. The ability to use more advanced methods of OCT imaging as a guide to potentially enhance lower quality OCT images can potentially allow for greater access to high end ophthalmic screening. By registering the high-resolution OCT images to their pairs of lower quality the training of a U-net model can be facilitated, which will be able to significantly improve the output of more accessible but less reliable imaging devices.

	Time-Domain OCT (TD-OCT)	Spectral-Domain OCT (SD-OCT)	Swept-Source OCT (SS-OCT)
Light Source	Broadband low-coherence source	Broadband low-coherence source	Narrowband tunable laser (swept source)
Typical Wavelength	~810nm	~840nm	~1050nm
Reference Mirror	Moving	Fixed	Fixed
Detector Type	Single photodetector	Spectrometer	Single high-speed photodetector
Scan Speed	Slow (~400 A-scans/sec)	Fast (~20.000–70.000 A-scans/sec)	Very fast (100.000 – 400.000 A-scans/sec)
Depth Resolution	Lower (~10μm)	Higher (~5–7 μm)	Comparable or higher (~5μm)
Imaging Depth	Limited	Moderate	High (penetrates deeper into choroid/sclera)
Interference Capture	One depth at a time	All depths simultaneously via spectral analysis	All depths simultaneously via wavelength sweep
Advantage	Original method; lower cost	High speed; high resolution;	Deep imaging; high speed;

Table 1
Concise comparative summary of different Optical Coherence Tomography modalities

2.3 Image Registration

2.3.1 Classification and Fundamentals of Image Registration

Image registration, within the field of medical imaging, refers to the process of establishing precise spatial correspondences between images. This task enables accurate alignment and facilitates the transfer of essential details and information across various captures of the same subject. [19] This correspondence not only allows for the ability to compare anatomical structures to the maximal possible precision, but can also facilitate computer vision or machine learning tasks such as segmentation or in our case, super-resolution.

Image registration can be categorized in various types depending on the source of the images and the modality used to obtain them. Unimodal registration describes the alignment of images obtained from a singular imaging modality, for example, multiple scans conducted using the same OCT) device. This type of registration is crucial in clinical settings for several reasons: it allows monitoring of disease progression over time, supports atlas-based segmentation processes, improves overall image quality through techniques such as image averaging, and provides essential guidance during intra-operative procedures. The process is often straightforward and algorithmically efficient due to the consistency in signal intensity and characteristics across images taken from the same device.

On the other hand, multimodal registration involves aligning images derived from different imaging modalities. These modalities might have fundamentally distinct operational principles, such as CT and MRI, or even different OCT devices that present substantial variations in feature

intensity and image quality. The inherent variability of multimodal images, primarily due to different physical properties and imaging mechanisms, introduces considerable complexity in accurately registering these images. Anatomical structures can exhibit markedly different appearances across modalities, thus requiring more advanced algorithms and methodologies to achieve effective alignment.

Despite its increased complexity, multimodal registration yields significant clinical and research benefits. It enables the integration of complementary information derived from multiple imaging techniques, each specialized in capturing different tissue characteristics—CT excels in depicting bone structures and dense tissues, while MRI provides superior visualization of soft tissues. This comprehensive approach enhances the diagnostic process and improves the efficacy of image-guided interventions. Moreover, multimodal registration holds the potential to facilitate the integration and interchangeability between sophisticated and simpler imaging devices, thereby democratizing access to advanced clinical information and decision-making tools. Harnessing and expanding upon this potential is a central focus of this thesis. [20, 21]

Image registration can also be subdivided based on who the subject of the images is. Intrapatient registration refers to registering images of the same person taken either from the same modality across time or across different modalities. Interpatient registration describes the process of aligning medical images between two or more different individuals. This endeavor can be highly challenging because of slight variations in normal anatomy across patients but yields high value in designing a standard frame of reference and in various research projects. [22]

A very simplistic breakdown of the image registration process of a pair of images can be described as follows: an image is assigned as fixed and its counterpart as moving. Next step in the sequence is to apply a transformation on the moving image to align it with the fixed image. Transformations can be divided into different types based on their characteristics:

i. *Rigid*: a fundamental transformation technique. If a rigid transformation is applied, distances and angles between different elements of the image remain unchanged. This approach is useful when the deformities between fixed and moving image are minimal and it is of utmost importance to maintain the existing anatomy without any deviation. Essentially, changes are applied only in terms of movement or rotation.
ii. *Affine*: is a more versatile technique that can apply the changes of a rigid transformation but also correct for alterations in scaling and shearing. It is very useful in medical imaging where malleable tissues can appear different across modalities and in other computer vision tasks.
iii. *Projective*: similar in logic, can apply all the changes that affine transformation can implement but also accounts for significant changes in depth variations and perspectives. More useful in non-medical applications, such as panoramic image stitching.
iv. *Non-linear deformations*: this approach is recruited when none of the aforementioned transformations are suitable to implement the changes needed for an accurate registration. Each pixel can be warped individually, allowing for maximum leniency. [23-25]

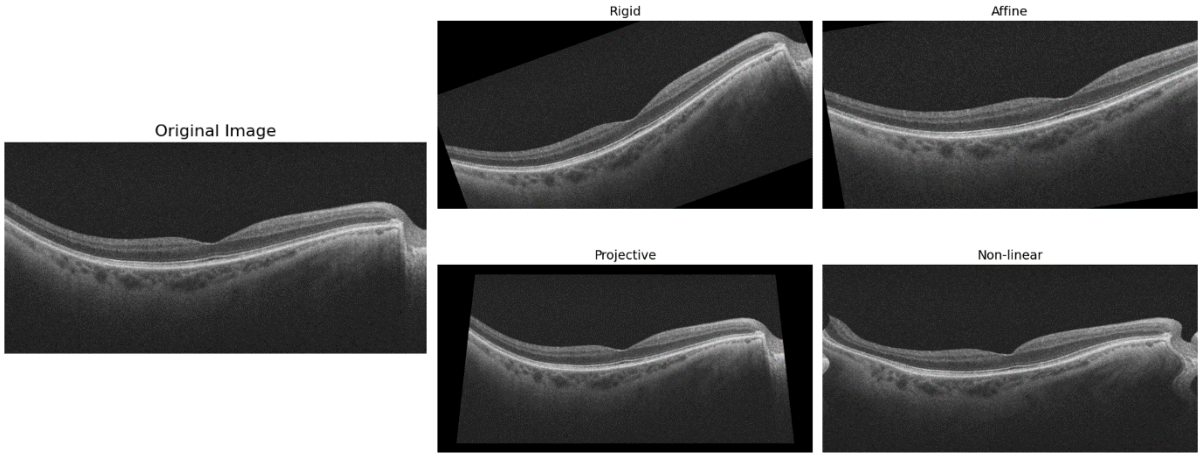

Figure 5
a) Transformation Examples. Original retinal image (left) and corresponding transformations using different registration techniques (right). Rigid transformation preserves shape and size. Affine allows for rotation, translation, scaling, and shearing, resulting in moderate distortion. Projective transformation simulates perspective distortions as seen from different viewpoints. Non-linear transformation models complex local deformations, such as tissue distortion or warping

Figure 5 acts as a sample of how different transformation approaches yield significantly different results and indicates how one must choose what technique they will employ based on the task at hand. Specifically, the rigid transformation uses a 2×3 matrix to rotate the image by 20 degrees and translate it by Δx = -26.68 and Δy = 210.55, while preserving shape and size. The transformation matrix $\boldsymbol{T}$ was applied to each image point by converting pixel coordinates [x, y] into homogenous form $[x, y, 1]^{\boldsymbol{T}}$ and computing the new location $[x', y']^{\boldsymbol{T}}$. The retention of shape and size is achieved by having the transformation matrix in the form of:

$$\boldsymbol{T} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & tx \\ \sin(\theta) & \cos(\theta) & ty \end{bmatrix}$$

Affine transformation also applies a 2x3 matrix, but can introduce scaling and shear. This allows for the change of angles and shapes, but parallel lines always remain parallel. Briefly, projective transformation changes the image and mimics the effect of looking at it from a different angle and non-linear transformation, bends and wraps the image in a wave like pattern without the use of a transformation matrix. In this case the formula was $x' = x + \Delta x(y)$, where $\Delta x(y) = 20\sin\left(\frac{2\pi y}{120}\right)$.

2.3.2 Classic Approaches to Image Registration

The fundamental logic behind image registration is to locate the areas depicting the same structure in different images and achieve a transformation that successfully aligns these

corresponding regions. The approach to achieving this task has revolved around two main characteristics of the images, intensity-based information and feature-based representation.

Intensity-based image registration methods align images by directly comparing their raw pixel values, eliminating the need for feature extraction or labelling of anatomy by experts. By relying solely on pixel intensities, these methods offer a straightforward approach, beneficial for scenarios where expert annotations are impractical or unavailable. The effective application of this approach relies heavily on two critical underlying assumptions. Firstly, the independence of pixel-to-pixel intensity values; specifically, each pixel's intensity contributes independently when computing the loss function that guides the optimal transformation. Consequently, these methods inherently disregard spatial relationships, such as structural continuity or contextual anatomical information and can thus limit the accuracy of alignment, particularly in cases of subtle anatomical features or complex structures. Secondly, the stationarity of intensity relationships between pixels is assumed; a consistent intensity correspondence between the two images throughout the entire imaging field is presumed. While global intensity discrepancies—such as overall brightness or contrast differences—can generally be managed by these methods, spatially varying variations in brightness or contrast can significantly challenge this assumption. Localized intensity distortions, commonly arising from imaging artifacts, uneven illumination, or scanner-related inconsistencies, may lead to misalignments. The presence of noise or localized artifacts further exacerbates this issue, making the optimization landscape more complex and prone to local minima, thereby compromising the reliability of registration results. Most commonly used techniques employed to achieve this kind of registration include sum-of-squared-differences (SSD), correlation coefficient (CC) and mutual information (MI). All of the aforementioned share the common goal of trying to minimize the total difference of pixel intensities or maximizing similarity when aligning two images. It is therefore easily inferred that such approaches can be less robust and reliable when registering images with varying spatial intensity distortions. [26, 27]

Feature-based image registration involves feature detection and description. Immediately, the difference lies on the fact that spatial relationships between pixels is accounted for. However, the processing needed is significantly more challenging than intensity-based image registration. Features that can be employed are lines, polygons, contours and more usually, due to the relative ease of description points. Keypoints must be detected and described, which translates to extracted as a distinctive area of the image and represented with no change when it comes to image deformation in any aspect respectively. The algorithms developed to achieve these tasks are often referred to as *detectors* and *descriptors*. The goal for a good detector is to be able to identify stable and distinct regions of the images despite potential transformations. [28, 29]

Most notable approaches to this task are Features from Accelerated Segment Test (FAST), the Scale Invariant Feature Transform Keypoint (SIFT), Binary Robust Independent Elementary Features (BRIEF) and Oriented FAST and Rotated BRIEF Features (ORB).

- FAST is a detector proposed by Rosten and Drummond, which aims to identify corners by examining the intensity of the neighboring pixels. The candidate pixel is identified if there exists a contiguous arc of pixels around said point which are significantly brighter or darker than the candidate’s intensity plus a threshold value.

It easily customizable as the radius of said circle can be set arbitrarily and efficient as only a subset of pixels in the circle is analyzed each time. In its simplicity however lies its disadvantage, as it can be less robust than similar models when faced with changes in viewpoints or other transformations. [30]

- SIFT is a four step algorithm proposed by Lowe et al.. Initially, a detection of scale-space extrema takes place, implemented by using a Difference-of-Gaussian (DoG) function. Its target is to identify keypoints which exist despite scale or orientation changes. Then those candidate points are localized and selected based on measures of their stability, i.e. points with low contrast are rejected as they are sensitive to noise. Then, one or more orientations are assigned to every keypoint location. This step ensures that all future operation will be performed on data that is transformed relative to assigned orientation, scale and location, rendering the results invariant to said changes. Lastly, local image gradients are estimated for the areas around the keypoint in the selected scale. This pipeline ensures that computationally expensive operations, like keypoint description, happen downstream, after a significant number of keypoints are discarded.[31]
- BRIEF is a descriptor proposed by Calonder et al.. It is designed with increasing speed of feature extraction and description in mind. Specifically, BRIEF performs comparisons in pixel intensity within a predetermined patch. The comparison yields a result of 1 if the first pixel is brighter and 0 otherwise. This allows for the formation of a string after a concatenation of all comparisons, which acts as a signature around a specified keypoint. Matching of descriptors can therefore achieved very quickly. However, BRIEF is not immune to geometric transformations and is heavily reliant on intensity, rendering it more sensitive to noise and lighting changes.[32]
- ORB is a combination of the FAST detector and an improved BRIEF descriptor. Proposed by Rublee et al., ORB is a keypoint descriptor that utilizes the advantages of the previously defined approaches. It utilizes FAST to define potential keypoints but also applies the intensity centroid method to estimate those keypoints orientation. After this processing it uses BRIEF to quickly assign a binary identity to each keypoint, with the added advantage of improving invariance to rotation.[33]

The aforementioned have proven effective in a wide range of image registration and computer vision tasks but nonetheless suffer limitations. Challenges revolve around limited robustness when changes to lighting or rotation are applied and significant computational workload. It is evident that an alternate approach needs to be adopted to tackle increasingly complex registration tasks.

2.3.3 Deep Learning Approaches to Image Registration

While classical approaches to image registration paved the way for contemporary endeavors, innovations in deep learning hold great promise to overcome obstacles that were

previously immovable. This is reflected by the trends of published literature regarding the topic in recent years as also depicted in Figure 6. This resurgence inspired the careful examination of a variety of Deep Learning models with a particular focus on their applicability on multimodal OCT registration, which led the selection of the model titled Reliable and Repeatable Detectors and Descriptors (R2D2). The following segment will briefly outline the most important contender models when the initial review was performed.

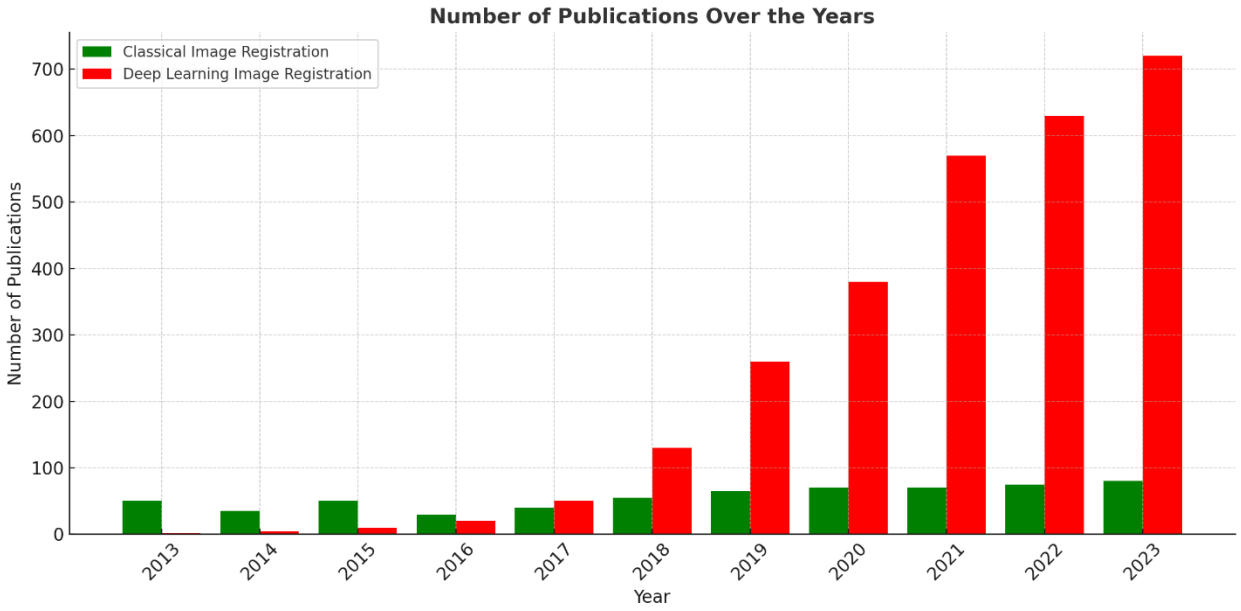

Figure 6
Representation of volume of publications in Classical versus Deep Learning image registration.

Voxelmorph was proposed by Balakrishnan et al. and aims to provide solutions for image registration tasks where deformation is required. It is mostly utilized in volumetric medical imaging, especially when attempting to register CT and MRI imaging of three dimensional (3D) structures. Voxelmorph employs a CNN to forego an objective function for each single pair of images and instead applies a generalizable mapping function that is procured through training on the particular subset of interest. This change in approach is reported to significantly diminish time demanded, even on a Central Processing Unit (CPU), while achieving the same or better results compared to other similar networks. However, there is two main reasons this approach may not be the optimal fit for the current project. Firstly, deformable changes, while potentially necessary to achieve registration between 3D volumes of CT and MRI, may significantly alter minute details that are of utmost importance in diagnosis and disease progression monitoring of retinal pathologies. Secondly, this thesis aims to register two dimensional (2D) images, a design which was not the primary drive in developing Voxelmorph. [34]

SuperPoint was proposed by DeTone et al. and presents a selfsupervised approach to identifying points of interest in 2D images. The focus of this effort is to negate the need of expert labeling of points of interest in the training dataset. This was achieved by first training the model on millions of synthetic images that portrayed simple geometric shapes and thus had no ambiguity

on what and where the keypoints were. This first iteration was named MagicPoint and performed well in real life situations, but trailed behind classical feature extraction algorithms. To address this issue the authors performed arbitrary homographic transformations on real life images and averaged the detector's response. Therefore, repeatability of proposed checkpoints was significantly increased and registration accuracy outperformed classical approaches. SuperGlue and later LightGlue are neural networks that were developed later to facilitate easier matching of the SuperPoint derived keypoints. Despite the promising results, the limited literature of the implementation of the aforementioned approaches in ophthalmic medical imaging, combined with the significant computational expense the need to train two separate networks –feature extractor and feature matcher- brings, render SuperPoint as not an ideal option for OCT keypoint extraction. [35-37]

Universal Correspondence Network (UCN) by Choy et al. is a fully CNN optimized for learning accurate visual correspondences. It focuses on identifying both geometric and semantic correspondences. Essentially, it aims to establish correspondences based on different viewpoints of the same still image and also correspondences of keypoints that are similar across different instances. It performs better than traditional feature extractors by actively looking for examples it is getting wrong during the training phase, a process referred to as hard negative mining. This is achieved because of previously annotated ground truth data. Overall, it is a fast and highly reliable model but it requires a supervised setting of learning and also can be computationally expensive because of its dense feature extraction and hard negative mining. [38]

Finally, Learned Invariant Feature Transform (LIFT), proposed by Moo Yi et al., is a deep learning pipeline that combines in a single model the full point handling array, that is detection, orientation estimation and feature description. Its novelty lies on the fact that while previous models tackled each of these problems individually, LIFT proposes to learn the aforementioned through a single CNN, which can facilitate better performance through the simultaneous optimization for all three elements. However, LIFT uses Structure-from-Motion (SfM) to accomplish its supervised learning. SfM identifies points of a 3D object that are consistent from different viewing angles and deems them important in the training process. This is not compatible with our task, since the OCT images are 2D slices of the same retina with no 3D viewpoint variation that is essential for SfM.[39]

In conclusion, classical and deep learning approaches follow a variety of techniques to achieve tasks in computer vision, medical imaging and other domains where image registration is essential. It is also evident, that each is more suited in certain tasks than others. Of utmost importance is to identify which model offers the optimal approach for achieving intrapatient multimodal OCT registration. The required characteristics include ability to drive affine transformations, 2D integration in training pipeline, computationally inexpensive and ideally previous implementation in similar tasks to assure effectiveness in OCT images. These conditions are most closely fulfilled by the model Reliable and Repeatable Detectors and Descriptors (R2D2)

2.4 Repeatable and Reliable Detector and Descriptor

2.4.1 Overview

R2D2, proposed by Revaud et al. is a detection and description approach that aims to simultaneously produce repeatable and reliable descriptors and keypoints. Its novelty lies in introducing an unsupervised training loss to learn a keypoint detector, a new loss to establish reliable local descriptors and a combined pipeline to produce both repeatable and reliable keypoints.

As mentioned in the previous chapters, most image registration and matching pipelines rely heavily on finding keypoints that are repeatable across variations of the same image, in order to establish them as guiding points during matching. While this quality is essential, solely focusing on this criterion can significantly limit the accuracy of the matching process, since repeatability does not guarantee that these keypoints will be useful or reliable. This is due to the fact that in many images where patterns repeat themselves, i.e. squares on a chess board or hyporeflective areas in OCT images, keypoints are consistently repeatable but suffer from decreased reliability due to their self-similarity. Essentially, since a keypoint is highly repeatable but lacks uniqueness, it can provide little to no valuable information when it comes to image alignment. This point can be further explained in figure 7.

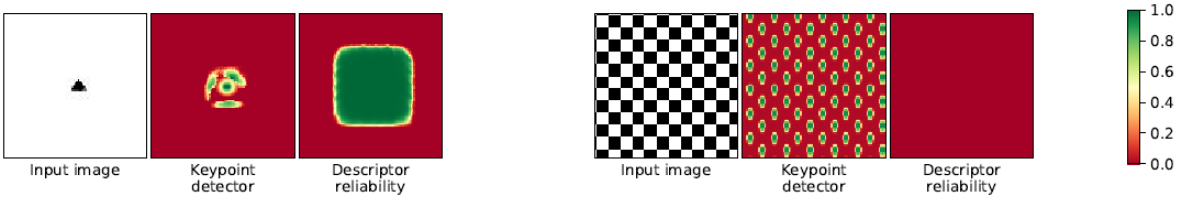

Figure 7
Demonstration of difference between repeatability and reliability. In the first input image (left) repeatable description can only be found in the center area where also reliability is high due to the uniqueness of the pattern. In the second input image (right), while repeatable keypoints can be found throughout the image but none of them are reliable due to inherent self-similarity.

Therefore, R2D2 provides the user with an important distinction on the extracted keypoints. Repeatable keypoints are procured, which determine their ability to be distinguishable features of the input image. This is paired with their reliability score that facilitates a more accurate matching process. In a nutshell, R2D2 can sort keypoints based on repeatability and reliability scores and provide the user with the optimal pathway for the matching endeavor.

Understanding this potential advantage acts as a motivation for R2D2 to jointly learn descriptor reliability and keypoint repeatability. The output to describe the two happens in two separate manners. First is through a score assigned to each keypoint that is the product of repeatability times reliability and secondly through maps that annotate distinctively the reliability and repeatability scores for every pixel in the image. This will be analyzed further later in this thesis. It is this integrated method that allows R2D2 to avoid highly repeatable but not distinctive regions for all its downstream tasks.[40]

2.4.2 Architecture and Loss functions

The R2D2 architecture adopts the backbone of a previously adopted L2-Net, which is also a compact descriptor extractor, with specific adjustments to better accommodate the output of both descriptors and reliability and repeatability maps after the input of a single image, irrespective of resolution.[41]

Specifically, while the L2-Net was originally designed to extract discriminative local descriptors from isolated small image patches of 32x32 and produce one descriptor per pass, R2D2 modifies the architecture to operate on whole images of any size or resolution and produces dense descriptors for every pixel and two confidence maps, one for repeatability and one for reliability. These adjustments facilitate the annotation of every pixel with a descriptor, a repeatability and a reliability score. [42]

The input image is first passed through a sequence of 3x3 convolution layers with batch normalization and Rectified Linear Unit (ReLu) activation functions through each step. The first correction R2D2 makes on the L2 backbone is to omit any downsampling in an effort to increase spatial resolution. Instead, it recruits dilated convolutions to enlarge the receptive field without compromising resolution or increasing kernel size. Dilated convolutions are able to provide this advantage by introducing gaps between kernel elements, effectively allowing the creation of relations between pixels that are further away than the kernel size might suggest. This is a critical step, as the target is to derive a descriptor for every single pixel, thus any downsampling or reduction of resolution is significantly opposed to the system philosophy.

Another adjustment is the replacement of the final 8x8 convolutional layer with a sequence of three 2x2 convolutional layers. This change can significantly reduce the number of weights without compromising accuracy.

The output of the final convolutional layer is a dense tensor of shape $\mathbf{H}_{\text{eight}}\text{x}\mathbf{W}_{\text{idth}}\text{x}128$. This tensor is the precursor for three of the network's output. First, l2 normalization is applied on to produce a descriptor of size $\mathbf{H}_{\text{eight}}\text{x}\mathbf{W}_{\text{idth}}\text{x}128$, able to be used for Euclidean-distance matching. Therefore, every single pixel has its own descriptor.

Simultaneously, the same precursor is passed through an elementwise square operation, which squares each feature value individually. By applying this operation, strong activations are amplified and noisy areas are suppressed. Then two parallel 1x1 convolutional layers are applied which reduce the tensor to single scalar value for each pixel and therefore produces two separate maps. These maps are spatially normalized using a softmax function.

Overall, this network is able to output a descriptor, as well as a repeatability and reliability score for every single pixel with a single forward pass on a full resolution image. This capacity can only be achieved through carefully designed and applied loss functions and training objectives.

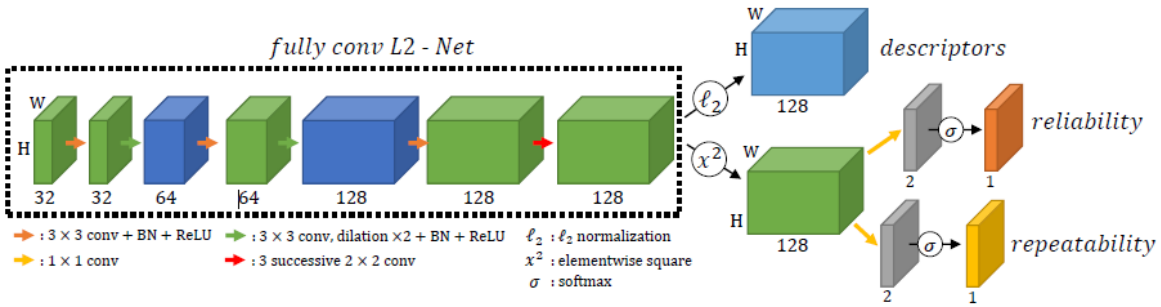

Figure 8
A representation of R2D2 architecture

R2D2 recruits a careful selection of loss functions and training objectives designed to optimize repeatability and reliability outcomes. They can be divided in two categories: repeatability losses and reliability losses. This distinction happens because repeatability and reliability are tow complementary aspects that must be predicted separately.

Learning repeatability revolves around the following functions. The authors claim that standard supervised training does not allow for the identification of novel detectors but rather copies existing ones and limits the results based on the shortcomings of each original method employed. This is why repeatability is a self-supervised task. This is achieved by applying the idea of maximization of cosine similarity between the repeatability maps of the two images over many small patches to avoid assumptions of zero occlusions or warp artifacts. Boiled down the network is encouraged to produce similar repeatability responses at corresponding locations with a loss that promotes the covariance between the repeatability maps. Self-supervision is achieved through the process of synthesizing a new image based on the existing one and thus already knowing the required transformation to fully align the two. This Cosim loss can be described as:

$$L_{cosim}(I, I', U) = 1 - \frac{1}{|P|}\sum_{p \in P} cosim(S[p], S'_U[p])$$

where:

- I: original image
- I': transformed version of the image created after applying a known transformation
- U: a dense correspondence field, ground truth where U[i, j]=[i', j'] if i,y represents a pixel in I and i', j' represents the same pixel in I'
- S: a repeatability heatmap generated by the network for image I
- S': a repeatability heatmap generated by the network for image I'
- P: the set of all overlapping NxN patches in the image domain
- Cosim(a, b): standard cosine similarity between vectors a, b. Measures angular similarity and not magnitude. This helps lower the dissimilarity based on different intensity values in OCT imaging.

This ensures that all local maxima in S will correspond with the ones at S'$_U$. However, this loss function can easily be minimized by applying constant values to S and S'$_U$. This is combatted by also applying a second loss function that intends to maximize contrast between keypoint and surrounding pixel, or in other words peakiness. This is achieved through the following function:

$$L_{peaky}(I) = 1 - \frac{1}{|P|}\sum_{p\in P}\left(\max_{(i,j)\in p} S_{ij} - \operatorname*{mean}_{(i,j)\in p} S_{ij}\right)$$

Overall, the final repeatability loss is:

$$L_{rep}(I, I', U) = L_{cosim}(I, I', U) + \lambda\left(L_{peaky}(I) + L_{peaky}(I')\right)$$

where λ determines the weight of the peaky loss in the overall repeatability loss.

Learning reliability involves assigning a confidence value for each descriptor ranging from 0 to 1 or otherwise $R_{ij} \in [0, 1]$. Higher values indicate greater reliability or in other words discriminating ability. The basic logic behind this approach comprises of vector comparison between the images. To be more specific, each descriptor from the image I is compared to all descriptors from the image I'. Since the applied transformation is known, then the corresponding descriptor is also known. This comparison can yield a rank of all the candidates with the goal of the true correspondence been at the top of this list. To translate this into a function a global metric called Average Precision (AP) is recruited. AP is a ranking based evaluation metric that combines precision and recall to measure the quality of the ranked results. In order to introduce an optimizable AP that can significantly contribute in the training process, the authors used a differentiable approximation of the AP, annotated as fAP. [43]

Overall, the relative function is the following:

$$L_{AP} = \frac{1}{B}\sum_{ij}(1 - fAP(p_{ij}))$$

where:

- B: number of patches in the batch
- Σ_{ij}: the summation over all pixel locations
- p: the patch centered at pixel (i,j)

To make sure that areas that are less distinctive are omitted the function is further optimized to look like this:

$$L_{AP,R} = \frac{1}{B}\sum_{ij}(1 - fAP(p_{ij})R_{ij} + \kappa(1 - R_{ij}))$$

where:

- B: number of patches in the batch
- fAP: Differentiable Average Precision score for given patch
- R_{ij}: Reliability score predicted by the network
- κ: a hyperparameter [0,1] to indicate the minimum expected AP per patch.

A more intuitive translation of this formula that can aid in tis comprehension is that keypoints will receive a high reliability score when their Average Precision score is also high as determined by 1-fAP(pij)Rij. The hyperparameter κ allows the model to determine how strict it wants to be in terms of allowing less discriminatory areas to be included. When increasing the κ value, lower assigned reliability values penalize the function, and thus descriptors with greater AP values become more important in the training process. The authors report that a value of κ=0.5 returns satisfactory results in practice but fine-tuning may be beneficial in certain tasks. [40]

2.4.3 Inference and Existing Models

During the extraction process, R2d2 models aim to extract repeatable and reliable keypoints from full resolution images. Other than the steps described earlier, R2D2 aims for scale invariability of derived keypoints and this is why it is not limited to a single image resolution. Irrespective of the size of the input image, it is rescaled, starting from the maximum of 1024 pixels as defined by the picture's largest dimension and is progressively downsampled till it reaches 256 pixels. It is important to make the distinction that this downsampling only happens during the extraction phase of the algorithm and not the training, since – as mentioned before – any such processing would result in the inability to produce a descriptor for every single pixel. At each different scale, the model produces the three desired outputs. For the keypoints to be meaningful in matching tasks, sparseness is essential and it is achieved by choosing only local maxima from the repeatability maps by applying a non-maximum suppression. Only pixels with repeatability values greater than their surrounding ones are included. Also, thresholds for repeatability and reliability values are defined at the beginning of the extraction process. Authors suggest a 0.7 score for both as satisfactory. [40, 44]

Pre-existing models have been trained using three distinct datasets, in an effort to increase the range of applicability. Those datasets, similar to the models previously reported, do not contain medical images, but rather web images and the Aachen Day Night dataset. [45] Each of these datasets provided 4000 images. Although no medical images were included in the training process of these models, numerous iterations of the R2D2 in those tasks have been really promising. [46-52] Interestingly, many of these endeavors have taken place in the realm of OCT imaging.

Specifically, R2D2 has been used to provide the framework for successful registration of image fundus photos. The increasing prevalence of retinal pathology and the necessity to create a comprehensive pipeline to diagnose pathologies and assess disease progression is the driving force behind those approaches, similar to the motivation behind this thesis. The application of R2D2 as

described earlier, was rigorously evaluated on the Fundus Image Registration Dataset (FIRE). [53]This dataset provides image pairs taken at the Papageorgiou Hospital, Aristotle University of Thessaloniki, that can be divided into three distinct groups. The first, contains image pairs with large space overlaps and no significant anatomical differences. The second group contain pairs of images that have undergone significant affine or rigid transformations and also contain images of non-healthy fundi. Lastly, the third group contained pairs of images with very limited overlap, which resulted in limited shared anatomical points. Villar et al. trained the existing R2D2 model on a different fundus photo dataset. They generated images after applying transformations on the original dataset fundus images to create the pairs necessary for training. The new model was applied on the FIRE dataset and was evaluated based on the accuracy of the image registration after some processing steps, like the application of Random Sample Consensus (RANSAC) to exclude outlier keypoints. The accuracy was based on a registration score. This score was calculated through the following processes. Initially, an error in pixels was computed for each image pair based on ground truth correspondences. A threshold is defined and any registration that falls below this threshold is considered successful. The percentage of successful registrations is plotted by varying the threshold from 0 to 25 pixels and the derived area under the curve acts as the score. For the first group, the proposed model achieved a near perfect score of 0.9275, for the second group the performance dropped to 0.726 and finally the most challenging pairs reported a 0.352 score. Those scores show the potential of R2D2 to facilitate image registration with minimal to none expert input even on the most challenging cases.[54, 55]

The authors proceeded to implement in similar approach to achieve inter device OCT image registration. However, the rationale was significantly different to what this thesis proposes. Essentially, the fundus images that were derived from each OCT device were used for registration purposes. This acted as a preparatory step to determine the “slice” of OCT that needed to be selected from each image modality to allow for the optimal inter device image registration. After the slice selection, the authors proposed to use a layer instead of simple keypoints to guide the registrations process and specifically the innermost retinal layer, the ILM. This required the annotation of said layer to register two similar in quality OCT devices. [56]

This brief review of related work highlights the use of fundus images as a means to achieve image registration both for fundus and OCT scan registration task. This approach occurs with good reason, as fundus photographs provide high contrast views of retinal anatomical marks such as the defining vasculature or other distinctive anatomical locations. On the other hand, OCT captures cross sectional details of retinal microstructures with significantly diminished ability to identify features and edges, the hallmarks of feature-based image registration. Additionally, the presence of noise and device specific contrast patterns compared to uniform representation of fundus images make the task of OCT based image registration inherently more challenging.[57-59]

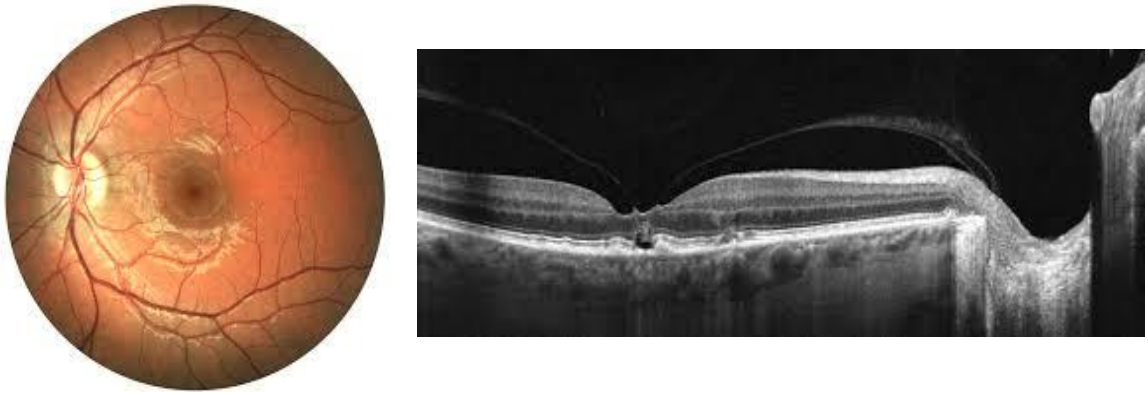

Figure 9
Comparison of fundus photography versus OCT images. The fundus photograph (left) provides high contrast landmarks that can easily act as features in order to drive keypoint detection. Those include vessels, the optic nerve head and other possible defining anatomical landmarks. On the other hand, the OCT image (right) lacks those distinctive features, hence increases the complexity of reliable and robust keypoint detection.

However, it is this complexity that renders the development of such a pipeline valuable. An algorithm capable of handling the intricacies that OCT scans yield can impact the field of ophthalmologic imaging in a manner which will allow increased accessibility and portability and subsequently earlier disease detection, optimal treatment outcomes and comprehensive monitoring.

Chapter 3

Reliable and Repeatable Detectors and Descriptors for Inter-Device OCT Alignment

3.1 Data Acquisition and Preprocessing

3.1.1 Device Overview

In our investigation towards multimodal image registration, two spectral domain OCT devices with distinctively different clinical utilities were employed. On one hand, the high-end Cirrus HD- OCT (Carl Zeiss Meditec, Dublin, CA) and the portable KUOS -O100 (Philophos, Daejeon, South Korea). Each device offers a unique set of advantages and limitations and provide an ideal scenario to evaluate the applicability of our pipeline. Both employ SD-OCT to provide their images. The spectral domain of the Cirrus HD-OCT is well documented, the KUOS -O100 was confirmed to be SD-OCT after personal communication with its developers. [60, 61]

The Cirrus HD-OCT device (Carl Zeiss Meditec, Dublin, CA), is a widely recognized OCT modality and commonly used by specialized ophthalmologic clinics for its high-quality derived images. It generates cross-sectional images of the retina that visualize intricate retinal structures and hence facilitate precise disease monitoring and clinical assessment. This system excels in contrast clarity and easy identification of retinal layers, delivering images with comparatively greater clarity and clinical information. Also, the Cirrus also benefits from a wider field of view, which allows the capture of a more expanded area of the retina. The pictures derived are 1920x991 in resolution and include a view of the fundus to annotate the slice used for the cross-sectional view.

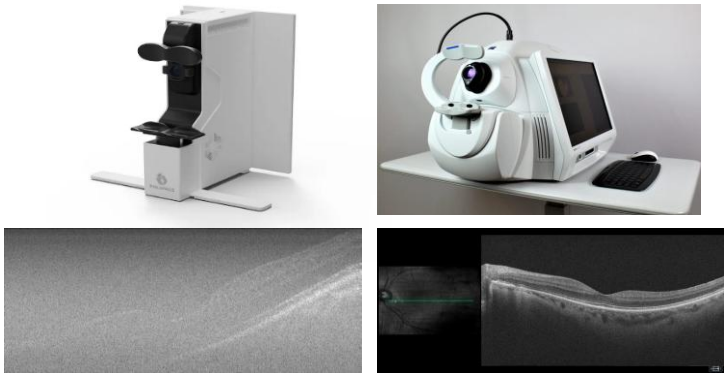

Figure 10
A depiction of the different OCT devices and the respective images they produce. On the top left, a portable OCT device which is significantly cheaper and easier to operate but suffers from low Signal to Noise ratio, small field of view and limited resolution. On the top right, a high-quality commercial OCT system can cause over $50,000 and can weigh over 50 pounds. Below each is the respective cross-section of the area they produce.

The KUOS -O100 (Philophos, Daejeon, South Korea) represents a different approach to OCT imaging that aims for its applicability outside the traditional clinical setting. It benefits from significant advantages in the areas of portability, accessibility, affordability and ease of operation. Those render this type of OCT modalities as perfect candidates for use in either resource limited or remote healthcare environments and can provide an exhilarating prospective of home OCT imaging. Images derived are 1024x512 in resolution. These benefits come at the cost of lower quality images that, while they may be useful in the diagnosis of easy to distinguish retinal pathologies, lack the needed output to discern minute retinal changes that often predispose to severe ocular illness. The output images frequently exhibit increased levels of noise, reduced contrast between retinal layers and more frequent artifacts. Overall, this OCT system is extremely efficient from a portability point of view but have limitations in terms of distinction of finer changes in retinal anatomy.

Those inherent disparities between the two devices acts both as motivation to develop an algorithm to achieve super resolution of the portable device images but also pose a great challenge in the registration process. The goal of this thesis is to achieve the production of detectors and descriptors that can overcome these significant challenges. This endpoint influenced the strategy of dataset construction, preprocessing and subsequent model training.

3.1.2 Data Acquisition and Initial Challenges

We collected retinal images to form a total of 84 pairs, each consisting of one image derived by the high-end and one by the portable OCT device. The images that formed the pairs were obtained within the same clinical session. The concurrent acquisition ensured minimal anatomical variation between paired scans; a step critical in the effort to achieve robust alignment.

The retinas scanned were carefully curated to include both healthy retinas and those diagnosed with AMD, hence both normal anatomy and one of the most common and debilitating retinal diseases could be represented in our cohort. Other retinal pathologies were excluded to streamline dataset consistency and to facilitate focused analysis specific to AMD-related changes. Pairings between the scans were derived by choosing the slice of both that most accurately depicted the foveal pit, hence ensuring close correlation between image points.

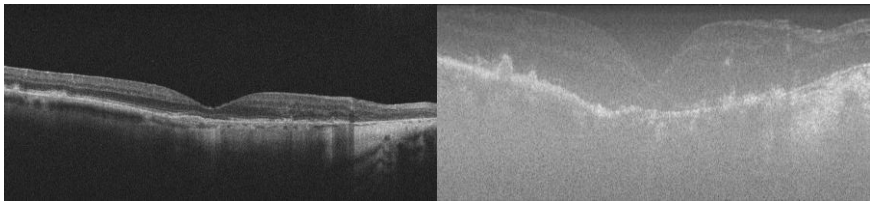

Figure 11
Example of an image pairing. On the left the picture derived form the HD-OCT device and on the right the picture derived from the pOCT device. Notice the significant differences in field of view, orientation and scaling between the images of the same retina. AMD related anatomical alterations can be seen.

The pairings immediately revealed significant challenges that could be encountered due to inherent differences between the images of the same retina. The commercial OCT device captured more expansive retinal regions while the portable OCT (pOCT) demonstrated comparatively smaller in scope retinal areas, often resulting in a field mismatch. Additionally, variations in rotation and inconsistent magnification characterized the pOCT images compared to the stationary Cirrus HD-OCT significantly increased the complexity of the task at hand.

Prior to moving forward with any preprocessing of the datasets, an existing model of the R2D2 was applied without any specific fine-tuning on training to assess the progress that can be made through specific alterations. After applying the extraction algorithm, no keypoints were derived. To further investigate the output, the repeatability and reliability maps were qualitatively evaluated. Figure 12 shows the maps for a specific image pair that were representative of all 84 image pairings.

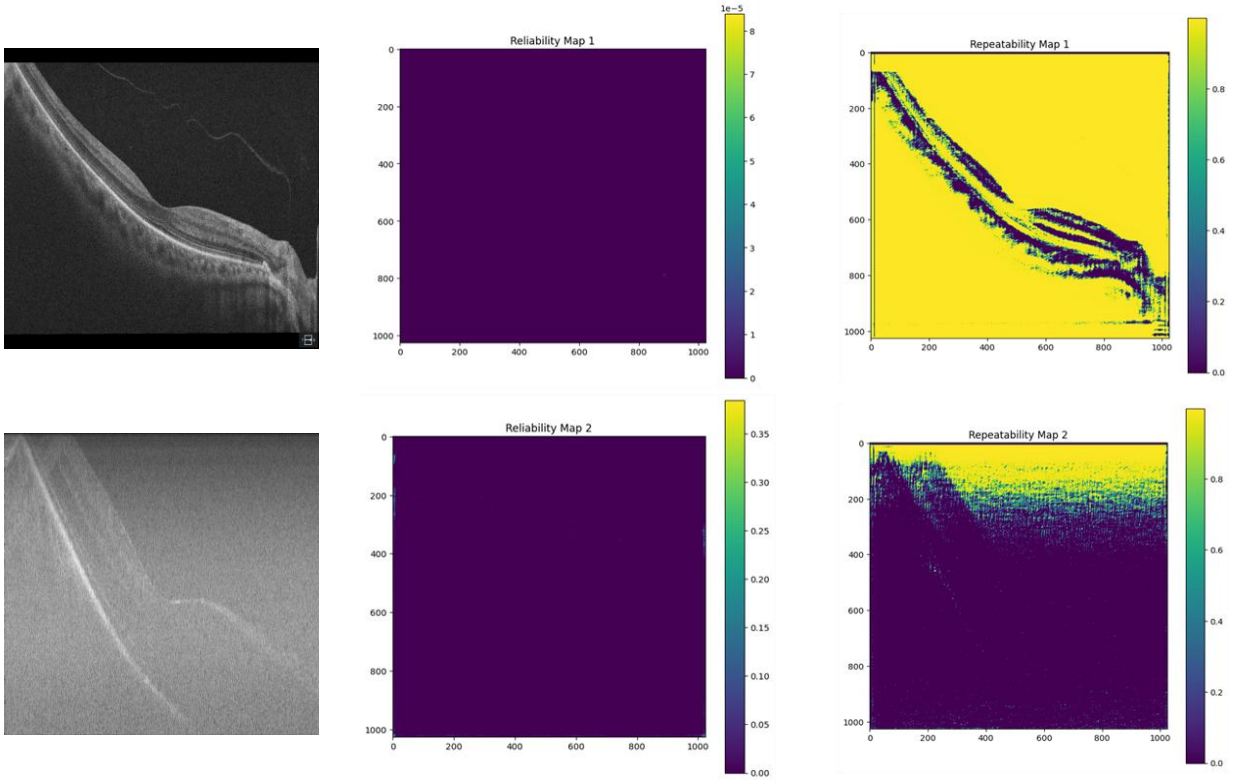

Figure 12
Representative reliability and repeatability maps after raw application of existing pretrained R2D2 model. Top row from left to right shows the image derived from the HD-OCT device, its reliability and then its repeatability map. Bottom row from left to right shows the image derived from the pOCT, its reliability and repeatability maps.

The reason for the inability to produce any keypoints can inferred by simply examining the reliability maps. Not a single pixel achieved a value significant enough to be depicted on the map, let alone to surpass the threshold which is set at 0.7 for both repeatability and reliability in the default model. However, repeatability maps can outline somehow accurately the anatomy of the

retina and individual layers in the HD-OCT data but this happens because low repeatability scores are assigned to retinal structures and higher repeatability scores on the hypo-reflective areas that annotate the hyaloid and areas posterior to the choroid. This highlights the importance of the premise of R2D2, that suggests that not all repeatable keypoints can be used effectively in alignment tasks. The task seems to be more challenging when it comes to the pOCT data as no concrete information regarding the retinal anatomy can be derived from the maps. Therefore, it is evident that further training and fine tuning must take place for the model to provide usable points of reference.

3.1.3 Approach towards a rough affine alignment.

A logical first approach to tackling most of the issues that arose during initial image pairing is to roughly align the images to overcome the inherent discrepancies between the images. While this seems counter-intuitive, as this is the goal of this project overall, the alignment does not have to be perfect and it will only be used in the training of the model to hopefully provide a solid foundation for clinical applicability without the need for further supervision. The choice of applying an affine transformation to achieve this goal was guided by specific considerations regarding anatomical integrity, complexity of matching elements and practical feasibility.

As described earlier in this thesis, an affine transformation is optimal for a number of reasons. Firstly, simple rigid transformations, which allow only rotation, are not suitable to achieve the aligning of the images due to the intricacies in scaling and skewing introduced by the pOCT images. Rigid transformations typically cannot accommodate these dimensional discrepancies, resulting in incomplete registrations that do not suffice. On the other hand, more complex non-linear transformations are computationally expensive and also pose significant clinical risks. Specifically, deformable changes especially in the RPE layer can significantly alter the retinal structure and in turn be misinterpreted by the clinician as erroneous pathological changes. Projective transformations can essentially accommodate for scaling and skewness differences, but their advantage is to correct for angle of perspective, a trait useful in panoramic image stitching, but of little value in medical image registration. Overall, affine transformations offer a solution that can compensate for differences in rotation, translation and scaling without compromising the anatomical validity of the retinal structures, as a key element of affine transformations is to keep parallel lines parallel.

Another compelling factor to employ affine transformations for this early alignment, is its simplicity. Essentially, the application of affine transformations requires the matching of at least three corresponding keypoints to produce the desired outcome. This translates, for our task, that if three anatomical points between the HD-OCT and pOCT images are correlated for each pair, we will have an initial alignment of the images that will play a pivotal role in model training later in the pipeline. [62]

3.1 Data Acquisition and Preprocessing

In order to increase the robustness of the matched keypoints, 5 ophthalmologists – 4 residents and a fellow with clear understanding of retinal OCT imaging – were tasked to deliberate in the choosing of these landmark points. Using a custom-developed, interactive Python based software application, the ophthalmologists consulted and chose in succession the points that they deemed as corresponding between the HD-OCT and pOCT image pairs. Mostly, the foveal pit was used as a guiding anatomical landmark. Other key landmarks that were used for pixel correspondence were areas were pathology could be discerned, i.e. drusen, or areas of thinning or thickening of retinal layers. All 5 ophthalmologists had to agree before proceeding to the next pair. No disagreements were recorded during this process. This expert-driven selection process ensured minimal observer variability and maximized the anatomical accuracy of alignment. The HD-OCT were chosen as the fixed image and the pOCT as the moving image. This involvement of ophthalmologists in the pretraining phase of the model development highlights once more the need for interdisciplinary approach to machine learning driven applications in the medical field. Figure 13 provides an example of images prior and after alignment.

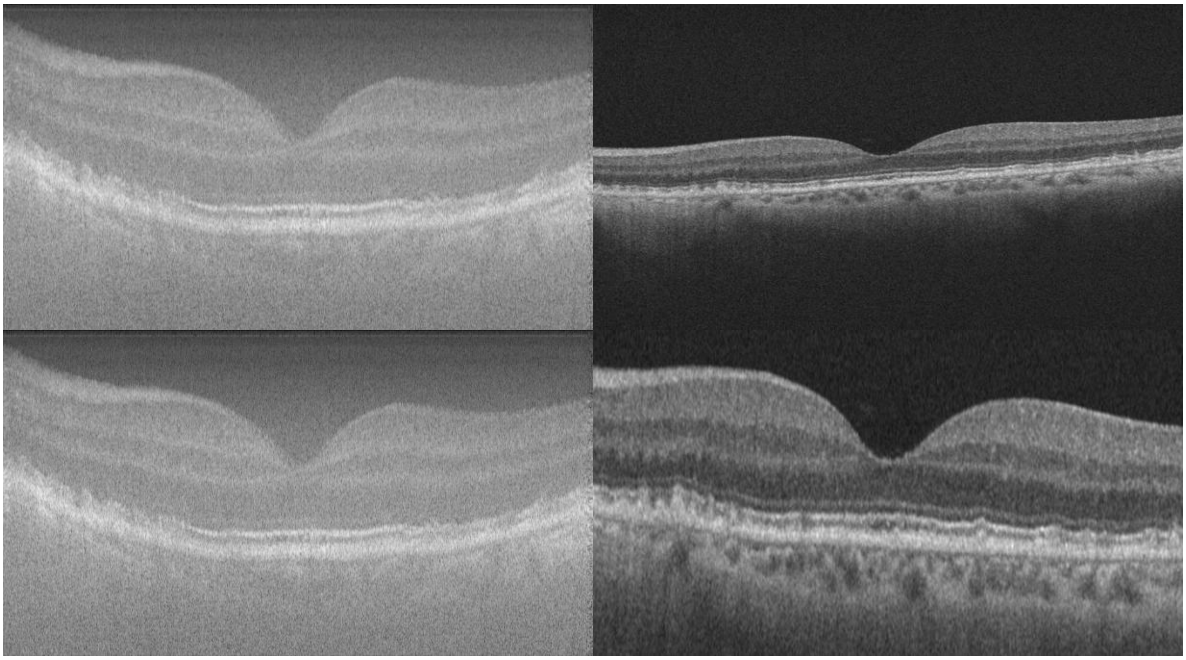

Figure 13
Example of rough affine transformation. On the left we can see the fixed p-OCT image. Top right is the original image provided by HD-OCT and bottom right is the resulting image after expert annotated keypoint affine transformation.

3.2 Dataset Creation

As discussed earlier, previous works that employed R2D2 utilized synthetic images generated from the original dataset to further train the preexisting models. This is done in an effort to finetune and optimize the derived model to accomplish the keypoint detection and description for the dataset it is intended. However, the challenges that the discrepancies between the image pairs pose may require an alternative approach.

The development of reliable datasets to finetune existing R2D2 models was a central component of this thesis' approach to multimodal OCT image registration. The quality and representativeness of training data is of paramount importance for the success of the model application. To this end, the creation of three separate yet interrelated datasets was performed in an effort to see which would yield greater results and potentially shed a light on how to optimally train R2D2 towards image registration between distinctly different images. The three datasets are the following:

- Crafted (C): This dataset was generated based on the premises outlined in the original R2D2 model and the related literature. It was synthetically generated to simulate realistic image transformations. All of the original images were resized to 256x128 to ensure computational affordability in the training stage. Each of the resulting images was treated as image I and synthetically generated images can be conceptualized as image I'. Transformations applied to the original images included a random combination of rotation, scaling, translation and perspective distortions at varying degrees of magnitude. Each reference image underwent three separate transformations, that resulted in three unique variants. The end result was a dataset that consisted of 504 image pairs and a dense optical flow field (aflow) that describe the pixel correspondence between I and I'. This setup is the input needed to train the existing R2D2 models. The main objective of this dataset is to enable the model to learn transformations in a controlled setting, where every pixel's transformation is precisely known. The controlled transformation parameters ensure that every pixel in the transformed image has a known correspondence in the original, enabling an accurate supervision that is not possible in real world datasets.

- Three-Point (3P): This dataset was generated in an effort to bridge the modality gap between different OCT acquisition systems. As mentioned previously, experts annotated 3 corresponding keypoints for every pair of images. This resulted in a rough affine transformation that aligned the retinal anatomy between images. The result of this process was 79 image pairs of the same retina, captured by a different modality and roughly aligned. Each pair consists of image A and B from the pOCT and HD-OCT respectively. In order to abide by the model's input demands however, an image pair and the dense optical field flow that connects them is required. To achieve that in a manner that has the potential to assist the model in learning modality-invariant descriptors, each image B was then treated as a source

image and subjected to the same transformation pipeline used in the C dataset. Specifically, seven random transformations were applied to each image B, generating transformed images B'. For every transformation, a corresponding dense optical flow field was computed, encoding pixel-level correspondences between B and B'. This allows for the creation of a dataset consisting of 553 image pairs and the aflow. These image pairs are A, B' and the B—>B' aflow. Although there is no precise supervision between A and B', this dataset organization encourages the model to consider A and B semantically equivalent, despite the inherent differences. This was an endeavor to allow the model to learn descriptors that are invariant not only to geometric distortions but also to cross device differences.

- Omni (O): this dataset was conceptualized in an attempt to explore whether combining the C and 3P datasets could lead to improved model performance in the task of multimodal OCT image registration. While C dataset offers precise pixel level supervision in a precisely self-supervised manner but in a fully synthetic but geometrically varied setting, the 3P dataset introduces real world modality differences with the best approach to overcome inter device dissimilarities but without the precision of pixel correspondence. Hence, they both share advantages and limitations. The O dataset was created to fuse the two, having as an objective to benefit from their potential. Half of its image pairs randomly came from the P dataset and the rest from the 3P dataset, resulting in a balanced dataset consisting of 500 image pairs in total.

3.3 Model Training

The fine-tuning experiments conducted leverage a pre-trained R2D2 model, which the authors refer to as fasterr2d2.pt. This model was extensively trained using diverse general imaging datasets, as described earlier. Notably, web images, Aachen Day/night synthetic pairs and their optical flow pixel correspondences between the original and the synthetically created image pairs. The pretrained model was fine-tuned specifically for multimodal retinal OCT images. Image pairs were consistently structured in directories, each containing two RGB retinal OCT images in .png format and their associated optical flow fields (aflow.pt). The image pair contained in each directory depended on the dataset used to further train the existing model.

Training parameters remained mostly consistent with what was used to train the original model to maximize compatibility and preserve learned generalizations. All three new models were derived after training for twenty epochs. The number was selected based on empirical evidence from previous iterations done by the author during the process of choosing the approach most suitable for the task. Other hyperparameters were set up as follows:

- Batch Size: set to 1 due to hardware limitations, ensuring memory efficiency during gradient computation and backpropagation

- Learning rate: 1e-4, a conservative learning rate used in the original training
- Weight decay: 5e-4, in alignment with the original training setting
- Optimizer: Adam, also used in the original training successfully
- Patch Size: 16, matching the pretrained model and ensuring compatibility[63]

The training script preserved the logic of the model provided by the R2D2 team. No changes were performed on the process during the three different iterations other than the location of input dataset directory and output model names and locations. The pretrained weights provided a stable initialization point. Average loss per epoch was logged for each iteration. During the fine-tuning process, occasional instances of Not a Number (Nan) loss values were encountered. Upon inspection, these anomalies were traced to image pairs where the synthetic transformation contained large empty or near black regions. When such areas were inadvertently sampled as image patches they produced degenerate input to the loss functions. Since these regions do not contribute meaningful supervision and can destabilize training, the adopted strategy was to automatically detect and skip such batches whenever a Nan loss was computed. Given that the overall dataset remained diverse and the majority of patches were unaffected, skipping did not negatively impact convergence or generalization. Upon training completion, three models were derived C.pt, 3P.pt and O.pt. Figure 14 shows the progression of the training loss during the training process for each model.

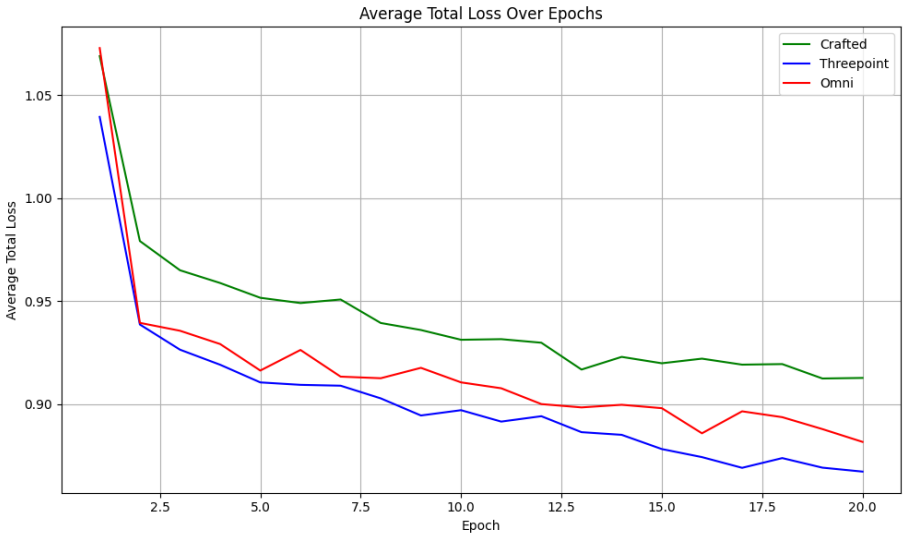

Figure 14
Graph demonstrating the progression of average training loss for each model across 20 epochs

All three models were initialized from the same checkpoint and trained under identical conditions. Among the three, the 3P model consistently achieved the lowest loss values across the training period, ending with the lowest average loss value, 0,8674. It also demonstrated the steepest

loss reduction in the earlier epochs. This early convergence suggests that the manually annotated affine pairs in the 3P dataset and the logic of considering image A and B as semantically equivalent, offered the foundation for keypoint detection and description. Additionally, 3P offered the lowest standard deviation of the three, indicating stable and consistent training progression with minimal oscillation. This is an encouraging first step in showing that semi-supervised learning is a viable approach to this endeavor.

The O model followed closely behind in terms of performance in training, converging to a final loss of 0.818 in the last epoch. It begun with the highest initial loss, which can be explained due to the greater heterogeneity of the training dataset. Nevertheless, it maintained a stable learning curve and was constantly in the middle of the three different models in terms of performance. This trajectory reenforces the value of combining diverse data sources.

Surprisingly, last in term of performance was the C model, even if by a close margin. The initial expectation that this model would be the best performer was based on the fact that the pairs of images that were input in the model were synthetic image data with precise control over transformations and accurate ground truth correspondences. Also, this was the rationale that was applied in relative work in the literature. Therefore, it was expected that the C dataset would provide the optimal training conditions for fine-tuning. On the contrary, the C model underperformed reaching a minimum loss of 0.9125 over the 20 epochs. These findings suggest that while synthetically crafted image pairs can support initial convergence, they may lack the variability necessary for robust keypoint descriptor learning in complex multimodal retinal data.

In summary, all three models demonstrated successful convergence under identical training conditions but also demonstrated differences in how each dataset shaped the learning process. The 3P model’s rapid and stable decline validates our underlying rationale that, once expert annotations are applied to bring paired images into alignment, each member of the pair can be treated as semantically equivalent for training purposes, at least in terms of loss behavior. The O model’s intermediate performance does not support the viewpoint that the combination of datasets can supply the model with both the precise geometric alignment of synthetic image pairs and the variability of multimodal OCT pairings to such an extent as to overcome annotated image pairs. Meanwhile, the lower unexpected ranking of the C model suggests that synthetic precision, while useful, may not be beneficial when it comes to OCT images due to the inherent noise and lack of distinctive features in those. To access practical utility however, we must examine the keypoints and descriptors derived by each network.

Chapter 4

Keypoint Extraction & Evaluation

To enable a direct and reliable comparison of the three detection models – C, 3P, O – all models were applied on the 84 intermodal image pairs. The primary objective is to extract and evaluate repeatable and reliable detectors and descriptors for each model under identical conditions and subsequently analyze their utility into downstream task of image registration. All models were initialized on the same pretrained R2D2 checkpoint under identical conditions and architecture, differing only in the datasets used. Hence, any observed differences in keypoint quality, quantity or score can be attributed solely on the effect of the training data itself. The inference setup was designed to neutralize external factors by keeping all parameters constant during the three runs.

Each model was applied to the entire dataset using a multi-scale keypoint extraction strategy, as described in the original R2D2 framework. At each scale level, the model produced pixelwise descriptors (128-dimensional), a repeatability heatmap, and a reliability heatmap. Local maxima in the repeatability map were selected using a 3×3 non-maximum suppression (NMS) operator. These candidate keypoints were then filtered based on their reliability and repeatability scores. The same thresholds were applied on all model runs. Due to the poor results of the original pretrained model, very conservative thresholds were set for both reliability and repeatability. Specifically, 0.1 score was set as a minimum requirement compared to 0.7 set by the original authors.[63]

The images were processed at multiple scales. They were progressively downscaled using a factor of $2^{0.25}$ until 256x128 resolution. At each scale keypoints were extracted, rescaled to the original image coordinates and stored. A score was assigned at each keypoint that was equal to the product of the repeatability and reliability scores and the top 5000 keypoints for each image were saved. The output of this process included the keypoint coordinates, the scale at which they were extracted, their scores and a dense descriptor for each. In a later process, repeatability and reliability maps were visualized for all images to better comprehend the distribution and quality of extracted keypoints.

4.1 Methods and Quantitative Results.

To evaluate the performance and robustness of the three trained R2D2-based models a comprehensive quantitative analysis pipeline was implemented to assess both cross-model and intra-model differences in keypoint detection and confidence. The objective was to systematically determine how each model performed on retinal image pairs acquired from our two distinct imaging modalities. As a foundational step, key metrics were calculated for each image and each model. These included the total number of detected keypoints, the mean keypoint score across all keypoints, which comprised of the product of reliability and repeatability score as described earlier, the standard deviation of descriptor scores, and the average score of the 10 keypoints that held the highest score. These metrics formed the basis for all subsequent statistical comparisons

and were selected to capture both the density and quality of detected features. To assess model performance independently of modality, the average score of the top 10 keypoints was compared across the three models. Intra-model consistency across modalities was also evaluated. For each model, HD_OCT and pOCT image performance was compared using the same three metrics (number of keypoints, average score of derived checkpoints and average score of the top 10 keypoints). For every image pair, the difference in scores between HD_OCT and pOCT images was calculated.

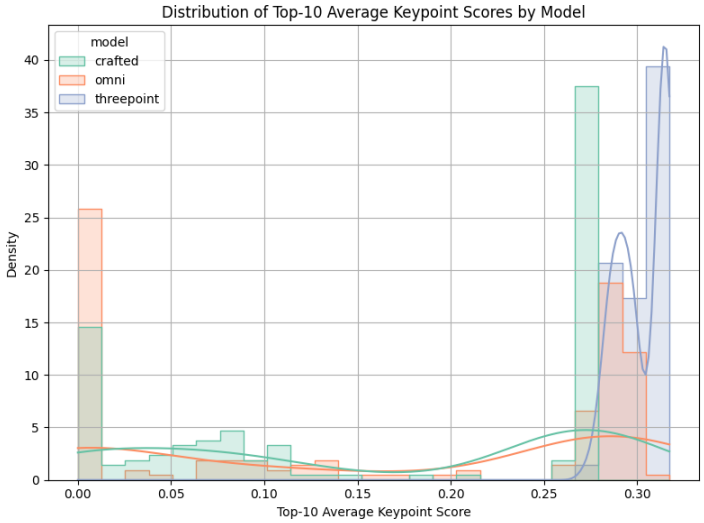

Figure 15.
Distribution of top-10 average keypoint scores for each model. The x-axis represents the average score of the top 10 keypoints detected in each image, and the y-axis indicates the density of these values across the dataset. 3P shows a tightly clustered distribution with consistently high scores, indicating reliable confidence across images. C displays a bimodal distribution, with a portion of detections near zero and another near 0.27, while O shows high variance with a notable concentration of scores near zero. KDE curves are overlaid to facilitate comparison of density trends.

Furthermore, descriptive assessment of keypoint extraction failure was conducted. Since a minimum of three keypoints is required to compute an affine transformation, any image for which a model produced three or fewer keypoints was considered a failure case for downstream registration. For each model and imaging modality (HD_OCT and pOCT), the total number of such failure cases was counted. These values were reported alongside the total number of evaluated images per category, and the corresponding percentage of failure cases was computed. This allowed a clear comparison of each model's capacity to produce geometrically useful keypoints across image types. No inferential statistics were applied in this context, as the goal was to characterize frequency and severity of insufficient keypoint extraction rather than test formal hypotheses. In addition to failure rate analysis, the variability of keypoint descriptor confidence scores was examined by comparing the standard deviation of scores produced per image across models. Higher standard deviation values reflect greater fluctuation in descriptor confidence within an image, whereas lower values suggest more uniform keypoint confidence. For each model, the standard deviation was computed for all images in the dataset, and the resulting distributions were compared pairwise between models. To analyze differences in keypoint detection behavior between AMD and normal cases, each image pair was annotated with a corresponding diagnostic label. For each model, independent comparisons were made between the

AMD and normal groups for four key metrics: mean score, top-10 average score, number of keypoints, and score variability. For all aforementioned comparisons, the Shapiro–Wilk test was first applied to assess the normality of the relevant distributions. Depending on the outcome, either a paired t-test or the non-parametric Wilcoxon signed-rank test was conducted for within-subject comparisons, and either an independent-samples t-test or a Mann–Whitney U test for between-group comparisons. A significance level of $\alpha = 0.05$ was used throughout, except in pairwise inter-model comparisons where Bonferroni correction was applied to account for multiple testing (adjusted $\alpha = 0.0167$).

Comparison of top-10 average keypoint scores across the three models was performed using pairwise Wilcoxon signed-rank tests with Bonferroni correction ($\alpha = 0.0167$). The difference between C and 3P was statistically significant ($p < 0.001$), as was the difference between 3P and O ($p < 0.001$). No significant difference was observed between C and O ($p = 0.108$). The median scores for C, 3P and O were 0.233, 0.308, 0.233 respectively. Average metric values for HD_OCT and pOCT images were also computed for each model. C extracted on average 11.3 keypoints for HD_OCT (mean top-10 score = 0.049, mean score = 0.040) and 5000.0 for pOCT (mean top-10 score = 0.273, mean score = 0.164). O detected 1.37 keypoints for HD_OCT (mean top-10 score = 0.038, mean score = 0.036) and 1092.89 keypoints for pOCT (mean top-10 score = 0.286, mean score = 0.101). 3P yielded 5000.0 keypoints for HD_OCT (mean top-10 score = 0.315, mean score = 0.249) and 4975.08 for pOCT (mean top-10 score = 0.291, mean score = 0.101). Intra-model modality comparisons were conducted for each model separately. For C, statistically significant differences were observed in top-10 scores (mean difference = –0.2235, $p < 0.001$), number of keypoints (–4988.67, $p < 0.001$), and mean scores (–0.1232, $p < 0.001$). O also showed significant reductions on HD_OCT in top-10 score (–0.2489, $p < 0.001$), number of keypoints (–1091.52, $p < 0.001$), and mean score (–0.0656, $p < 0.001$). In contrast, 3P demonstrated significantly higher performance on HD_OCT compared to pOCT in top-10 scores (mean difference = 0.0241, $p < 0.001$), number of keypoints (24.92, $p = 0.023$), and mean score (0.1482, $p < 0.001$). The failure rate, defined as the number of images with three or fewer detected keypoints, was also recorded. C failed in 57.1% of HD_OCT images and 0% of pOCT; O failed in 91.7% of HD_OCT and 0% of pOCT; 3P failed in 0% of both HD_OCT and pOCT. To assess variability in keypoint descriptor confidence, standard deviation of scores was compared between modalities within each model. C showed significantly lower variability in HD_OCT (mean difference = –0.0193, $p < 0.001$), as did O (–0.0551, $p < 0.001$) and 3P (–0.0261, $p < 0.001$). Inter-model comparisons of standard deviation of scores were also performed across all images. C was found to be significantly less variable than both O ($p = 0.0020$) and 3P ($p < 0.001$), while no significant difference in variability was observed between O and 3P ($p = 0.198$). Regarding AMD versus normal comparisons, no statistically significant differences were observed for any model across any metric. For the 3P model, the difference in mean score between AMD and normal cases was not significant ($p = 0.810$), nor was the difference in top-10 average score ($p = 0.643$), number of keypoints ($p = 0.896$), or score variability ($p = 0.481$). Similarly, for the O model, no significant differences were observed in mean score ($p = 0.902$), top-10 average score ($p = 0.765$), number of keypoints ($p = 0.863$), or score variability ($p = 0.637$). The C model also yielded nonsignificant results across all metrics: mean score ($p = 0.486$), top-10 average score ($p = 0.344$), number of keypoints ($p = 0.406$), and score variability ($p = 0.290$)

Number of Keypoints per Model x Modality			
Model	**Modality**	**Mean Number of Keypoints**	**Percentage (%)**
Crafted	HD	11.33	0.23
Crafted	p	5000	100%
Omni	HD	1.37	0.03
Omni	p	1093	21.86
Threepoint	HD	5000	100
Threepoint	p	4975	99.5

Table 2

Mean number of keypoints detected per model and modality, along with their percentage relative to the theoretical maximum (5000 keypoints). While 3P consistently reaches or nearly reaches this maximum across both modalities, C and O show severe degradation in HD_OCT images, producing less than 1% of possible keypoints on average.

The comparison of the top-10 average keypoint scores among the three fine-tuned models reveals meaningful information in model behavior and keypoint quality. The results indicate a clearly significant advantage of the 3P model over its two counterparts. No statistically significant differences are found between C and O. This quantitative finding validates the rationale used for training the 3P model and allows for similar future applications in the keypoint extraction pipeline. The top-10 average key score metric captures the model's ability to assign higher confidence to the most relevant or potentially more robust features in the image. It is noteworthy to mention that while the C model was expected to perform the best due to the similar logic of the training dataset to the one used in the original R2D2 paper, it did not manage the expected results. Additionally, the assumption that the O model could gain an advantage of both rationales – the precise pixel correspondence of the C dataset and the multimodal invariance the 3P model theoretically offered, did not materialize as it also underperformed quantitatively in this metric. This suggests that the integration of the two datasets in the O model potentially did not have a synergistic effect and indicates that the learning process in this model was more heavily influenced by the C rather than the 3P dataset. Also, no differences noted between AMD and normal groups across all models suggest that model behavior is consistent, at least in terms of quantitative results.

The averaged performance metrics across modalities further clarify the differential capabilities of the three models and reinforce the earlier statistical conclusions. The 3P model once again stands out, achieving not only the highest mean top-10 keypoint scores in both HD_OCT (0.315) and pOCT (0.291), but also maintaining a consistently high keypoint count near the maximum (5000 and 4975.08, respectively). These results suggest that 3P is not only capable of detecting keypoints, but also that these keypoints are assigned high confidence values. In contrast, the C and O models demonstrate a severe performance drop when transitioning from portable OCT images to high-definition OCT images, and admittedly puzzling result. While they extract more than 1000 keypoints on average in pOCT, their keypoint count plummets to just 11.3 and 1.37 on HD_OCT. A similar trend is evident in their mean top-10 scores, dropping from 0.273 in pOCT to just 0.049 in HD_OCT for C and 0.286 in pOCT to 0.049 in HD-OCT. These limitations further highlight the challenge of tackling multimodal OCT registration. Despite relatively good performance from all the models on the pOCT images, the low scores and very low number of

keypoints derived from C and O on the HD OCT images make the registration task using keypoints from these two models almost impossible. This realization is further supported by the fact that the repeatability and reliability thresholds set for keypoint extraction were already significantly lower compared to what the authors originally suggested -0.1 compared to 0.7- hence even these low standards could not be met by two out of the three trained models. Also, 3P's apparent success comes into question under the same rationale. The permissive thresholds have allowed for high detection counts, perhaps suggesting that very weak keypoints are extracted. This reasonable doubt can only be confirmed or rejected when performing a qualitative assessment or during the image registration procedure.

Finally, as mentioned earlier, at least 3 points are required to produce the matrix for the affine transformation needed to perform image registration. As such, any image that has less than three points annotated after the extraction process cannot be introduced in a registration pipeline. This limitation further highlights the shortcomings of C and O in this domain as the failed to overcome this barrier in 51.7% and 91.7% of HD-OCT images respectively. These failures occurred under very relaxed score thresholds, a condition that underscores their inability to be employed in the requested pipeline. On the contrary, 3P was able to produce close to the max number of possible keypoints in both modalities. These results decisively disqualify C and O from use in image registration.

Despite the 3P's model seeming superiority in across all metrics, its value as a meaningful keypoint extractor and subsequently its decisive addition in an automated or semi-automated image registration pipeline remains to be seen. High numerical scores and ability to produce consistently the necessary number of require keypoints are not sufficient on their own to guarantee anatomical relevance, especially in the context of retinal OCT images where spatial correspondence of keypoints to anatomical landmarks is vital for registration. It is still a possibility that the promising results thus far are a result of lenient thresholds or that the derived keypoints are anatomically uninformative. By that, it is implied that if keypoints describe areas in the hyaloid -anterior to the retina- or posterior to the choroid where signal intensity is almost nonexistent in both modalities, then they cannot be the guiding force in a registration task as the results will certainly be unreliable. To make the distinction between useful and unreliable model, a qualitative assessment of derived keypoints is necessary. This can be achieved by examining the output of repeatability and reliability maps versus the original images. In this manner, it can be decided if the keypoints derived truly align with relevant retinal landmarks or are instead the result of statistically inflated but functionally arbitrary detections.

4.2 Qualitative Evaluation

To qualitatively assess the anatomical relevance of derived keypoints by each model it is imperative to explore their distribution in the image and their relative repeatability and reliability scores by carefully examining the corresponding heatmaps. This analysis focuses on whether those

maps consistently outline retinal layers and have the ability to disregard areas with high noise or little to no relevant clinical information. To be more specific, areas anterior to the ILM or posterior of the choroid lack the ability to guide image registration as they do not have any localizable anatomical landmarks. This was also the limitation of the original pre-fine-tuned model, as mentioned earlier, which tended to assign the highest scores to homogenous areas, low information areas while suppressing key regions of retinal anatomy – thus undermining any cross-modal registration effort. (Fig. 12)

The rationale behind this qualitative evaluation approach was to systemically assess all the repeatability and reliability maps derived by each model. Initially, HD-OCT and pOCT images were evaluated independently to determine model performance on each set. Subsequently, maps of different modality pairs were compared. Pattern recognition was the main objective of this process. Within the scope of this process is to determine which model better assigns keypoints scores based on the criteria mentioned previously, if any model is superior in terms of providing a similar distribution of keypoints across modalities and also to determine the reason behind low reliability performance for the C and O group. The following section is dedicated to going over some representative such maps to further delve into how the models C, 3P and O function and if the quantitative results are also corroborated here or if new information becomes available that alters the rankings of the models.

4.2.1 HD-OCT repeatability and reliability heatmaps evaluation across models

The qualitative assessment of the repeatability and reliability maps revealed important distinctions in model behavior with special focus in their potential to emphasize anatomically and clinically relevant regions while suppressing background noise. The C model consistently demonstrated the most desirable properties in its repeatability maps. Areas within the retinal contour, that is between the ILM and Choroid layers, were densely populated with high scoring keypoints while areas of little to no anatomical relevance were correctly suppressed. This spatial selectivity can be critical as it decreases the likelihood of erroneous and arbitrary matching of areas lacking valuable spatial information. This exceptional performance creates a stark difference compared to the very poor results C had in its quantitative analysis. This further highlights how challenging the process of keypoint extraction and multimodal image registration can be as quantitative and qualitative results may be significantly different, hence calling for an inventive solution.

In contrast, the repeatability heatmaps by the 3P and O models demonstrated similar profile with lower discriminative ability. Although both models were capable of vaguely outlining the retinal contour, they frequently annotated high repeatability scores for pixels that fell outside the field of relevance. While this is not unexpected given the definitions provided earlier, where reliability is used as a metric to rule out repetitive areas, it is a significant disadvantage of these two models compared to C. Those results however fail to explain the disparity displayed earlier, where 3P achieved by far the greater performance in all areas when it came to HD-OCT images.

The answer to this paradox is given by careful examination of the reliability maps. Both C and O models had mostly completely empty reliability maps, indicating their weakness in finding and extracting reliable scores for their keypoints. This sparsity of any scores in reliability maps especially nullifies any exceptional repeatability performance and clearly explains why C and O failed to produce many times more than the 3 necessary keypoints to guide affine alignment. However, we must keep in mind the exceptional performance of C in repeatability and aim to take advantage of its annotations downstream in the pipeline.

One of the most challenging interpretations is that of the 3P reliability maps. While it justifies its excellent performance in our previous analysis, it underscores one of the model's limitations. High reliability scores were assigned to areas lacking anatomical marks or textual information of clinical interest. Conversely, areas which would ideally be marked with high scores, had almost always a zero or near zero score. This inverse scoring pattern undermines the superiority demonstrated earlier and renders the ranking of the trained models highly ambiguous. However, 3P tends to produce stark contrasts between highly reliable and non-reliable points right at the ILM and hyaloid border and at the RPE as well. This high gradient may be useful in determining the boundaries of the retinal contour and can perhaps be utilized downstream.

This model behavior remained consistent throughout the dataset despite the absence or not of pathology. This allows us to make generalized conclusions based on the patterns described above and also demonstrated in Figure 16. The C model stands out in its ability to localize keypoints rich in anatomical information with concurrent suppression of irrelevant areas or areas that would not be ideal to include in an OCT multimodal image registration task. However, its inability to assign high reliability scores in a similar manner significantly impairs its value in extracting the optimal descriptors. The 3P and O models exhibit more general and not specified repeatability maps, with O producing the least desired reliability maps, both explaining its low performance in quantitative analysis and also rendering it last in model performance. 3P also produces misattributed reliability maps, a fact that adds an additional challenge in robust registration.

Nevertheless, the models thus far show a complementary profile in terms of their keypoint extraction ability. C's excellent repeatability performance and 3P's reliability maps that can successfully showcase the retinal borders can be fused to produce a combined approach to the registration process.

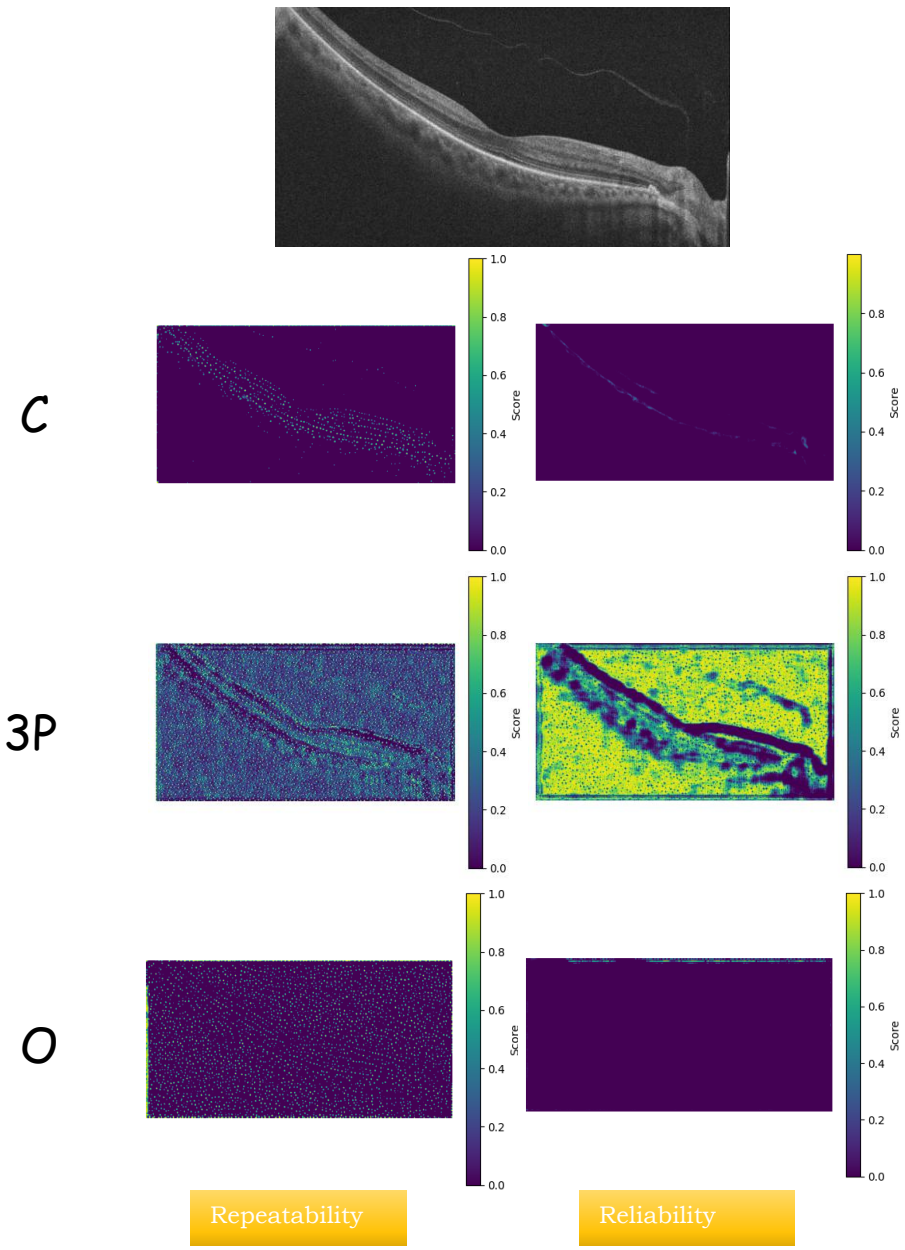

Figure 16

Repeatability (left column) and reliability (right column) heatmaps generated by the three evaluated models—Crafted (C), Three Point (3P), and Omni (O)—for a representative HD-OCT image (top). The C model produces a highly selective repeatability map which also" catches" the detached hyaloid contour, while suppressing irrelevant regions. However, its corresponding reliability map is nearly empty, explaining its limited keypoint utility despite high spatial precision. The 3P model displays a broader and noisier repeatability map and a reliability map with high scores in clinically irrelevant areas, failing to emphasize meaningful structures. The O model shows minimal output in both maps, reflecting its overall poor feature extraction performance in HD-OCT scans.

4.2.2 pOCT repeatability and reliability heatmaps evaluation across models

When examining pOCT images, model behavior exhibited similarities compared to its earlier output on HD-OCT images, however certain aspects shifted substantially. The C model continued its excellent outputs in repeatability maps, assigning high repeatability scores to areas corresponding to retinal structures. The difference lied on the model's ability to suppress areas of low clinical significance diminished and high repeatability scores were also assigned at the hyaloid and posterior to the choroid. Hence, this model's capacity of discerning between areas of retinal contour and noise did not have the same impressive results in this instance. Nevertheless, it is easy to locate the retinal layers when observing C's repeatability maps, despite its diminished performance in this front. Surprisingly, the reliability maps improved when examining the lower quality images of the portable OCT modality. In contrast to the sparsely populated maps produced previously, pOCT reliability maps included more areas within retinal layers but, similarly to the repeatability output, high scores in the choroidal region or the hyaloid limited the discerning potential of these keypoints.

The 3P model demonstrated the most invariable output between the two modalities, exhibiting similar performance and keypoint distribution in both. This is an expected outcome since it is the model that most "forced" the training algorithm to consider the two modalities equal and as such should produce invariant results, at least in theory. In actuality, the repeatability maps continued to capture retinal contour efficiently, with similar prowess in isolating retinal contour compared to other structures of the eye. Similarly, the reliability maps followed the same pattern; low emphasis on clinically relevant regions and high assigned scores in the rest with a strong contrast across the two. However, in this instance the margins were less clear compared to the HD-OCT output, a result that may be attributed to the high noise of the pOCT images.

The most notable improvement was observed in the output of the O model. While its repeatability maps remained broadly similar to those seen in HD-OCT, with diffuse and less selective keypoints, its reliability maps demonstrated a marked improvement. Compared to the almost non-existent assignment of scores in the HD-OCT images, the pOCT reliability maps accurately captured the spatial profile of the retinal contour, but also scored highly in regions outside of it. Among the three, O may have achieved its most balanced and modality-aligned scoring in the pOCT, perhaps suggesting a potential for modality specific optimization.

Overall, these results illustrate distinct model-specific behaviors with respect to modality. The 3P model exhibits the greatest consistency, maintaining similar performance across both HD-OCT and pOCT. The C model, while superior in HD-OCT due to its highly selective repeatability maps, experiences a slight decline in specificity on pOCT, though with improved descriptor reliability. Conversely, the O model shows limited utility on HD-OCT but a notable improvement in reliability scoring when applied to pOCT data. This shift may point to the potential of modality-specific optimization for enhancing registration performance in future work. Once more, no model clearly dominates in both repeatability and reliability fields, with each demonstrating significant advantages and drawbacks in extracting the ideal keypoints. This raises the question whether each can be used on its own to provide the optimal image registration pipeline or if a fusion approach

may benefit from each model's strengths. Figure 17 presents a pOCT image with representative for the group repeatability and reliability maps for each model.

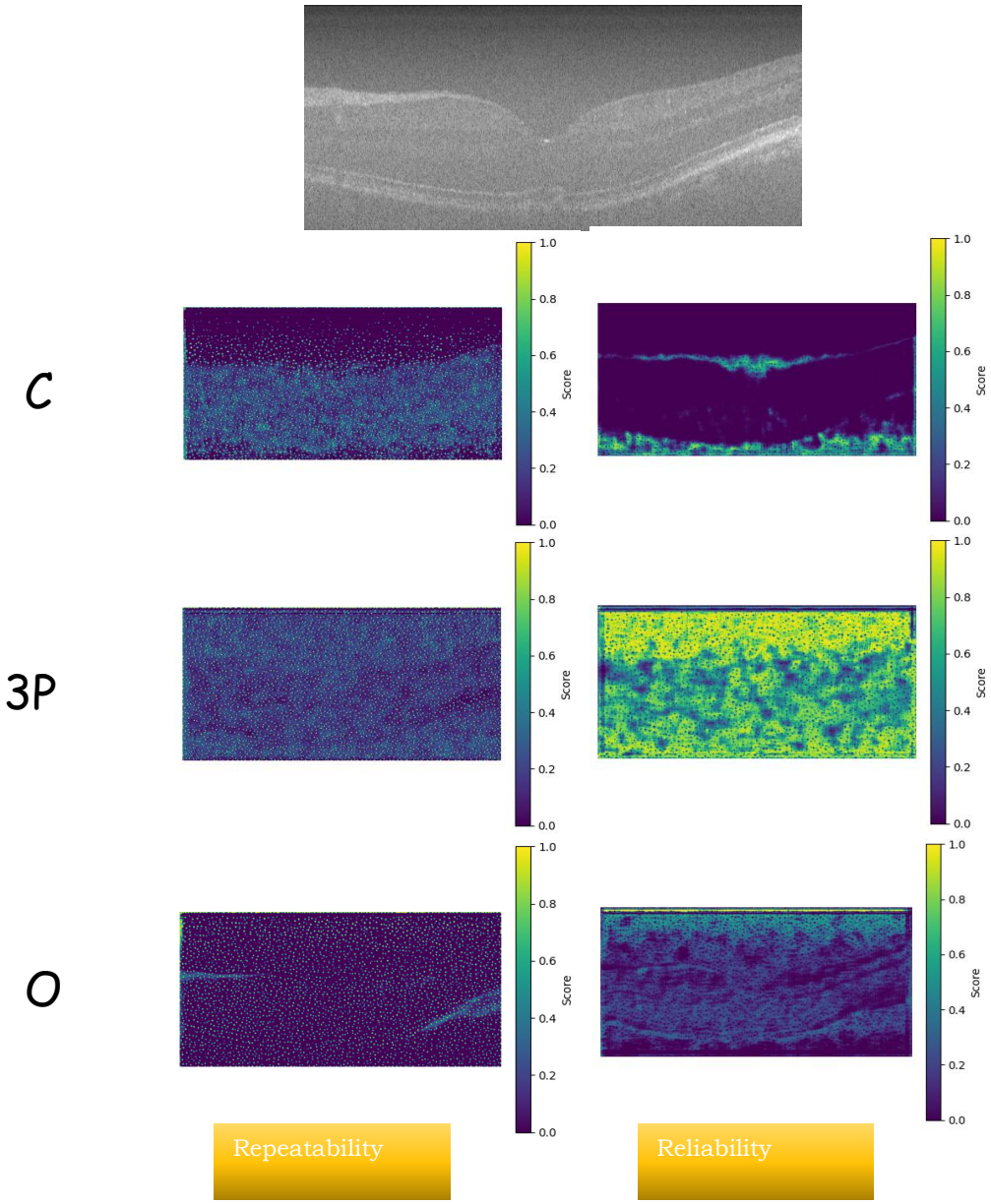

Figure 17.

Representative repeatability (left column) and reliability (right column) heatmaps from a pOCT image for the three evaluated models: Crafted(C), Three Point (3P), and Omni (O). The C model displays acceptable repeatability along the retinal contour but fails to clearly demarcate it from areas of low clinical significance. Reliability output improves compared to HD-OCT, with more populated maps. The 3P model exhibits consistent behavior with similar repeatability profiles to those of the HD-OCT images and a reverse to the desired output in the reliability maps with less clear distinction of the retinal contour borders. The O model retains diffuse repeatability but shows the most notable improvement in reliability maps, with scores that better align with the retinal anatomy. This convergence between detection and confidence suggests improved coherence in keypoint selection for the O model in pOCT images.

4.2.3 Intramodality keypoint correspondence evaluation

Having evaluated each model's repeatability and reliability performance within individual modalities, the next vital step to assess was to examine how consistently each model identifies corresponding anatomical features across modalities. This step aims to address a critical requirement for multimodal image registration, which is that the extracted keypoints must not only be reliable and repeatable but also consistent across domains. If a model assigns its top scoring keypoints to anatomically analogous areas across modalities, it demonstrates potential to successfully guide the image registration objective. On the other hand, if said model scatters the keypoints in distinct areas across the modalities, then no meaningful correspondence can be achieved despite any optimization processes.

Unlike sections 4.2.1 and 4.2.2, which focused on keypoint metric evaluation and qualitative assessment of heatmaps for each individual model on any one given modality at a time, this chapter concentrates on a sparse set of highest performing keypoints selected for downstream use. This approach can illuminate the process of how a model will perform in practical applications where only a specific subset of keypoints can be used to define the transformation. Given the high failure rate the O model returned and to ensure a meaningful comparison, 5 pairs were the O model produced more than the minimum number of required keypoints was produced, were selected. The keypoints were plotted on top of the images and a color grading was applied indicating the highest to lowest scoring keypoint based on color.

The results of the comparison further highlight the significant intricacies and challenges that inhibit a straightforward approach to multimodal image registration. While each, image may yield repeatable and reliable keypoints, if those do not correspond across modalities, there is almost no viable solution to define an accurate transformation matrix. This also applies in our case. The C model's top 100 performers annotate a very narrow retinal area close to the RPE or Optic Nerve Head in most cases, when applied on HD-OCT images, while the keypoints in the pOCT counterpart are more widely scattered and seldom correspond in their location. 3P almost completely misses the retinal contour in both instances. While the keypoint distribution is similar across modalities, it does not reflect meaningful structures and it is highly doubtful whether they can help navigate an accurate affine transformation. The O model rarely achieves greater than three keypoint in the OCT images and when it does, they are representative of a small area that almost never corresponds with keypoints of the pOCT images.

Taking into account all of the aforementioned data, it is safe to draw specific conclusions. Firstly, the consideration of two roughly affined images as equal in terms of model training is a valid approach that can be corroborated both quantitatively and qualitatively and can be applied in future keypoint extraction endeavors. Additionally, quantitatively discouraging results do not necessarily imply poor overall model performance as certain advantages of the lower scoring C and O models were discovered when carefully examining the distribution of the heatmaps. Finally, traditional approaches to image registration in our instance may have a more challenging application. Alternative approaches leveraging the advantages of each model may be warranted.

4.2 Qualitative Evaluation

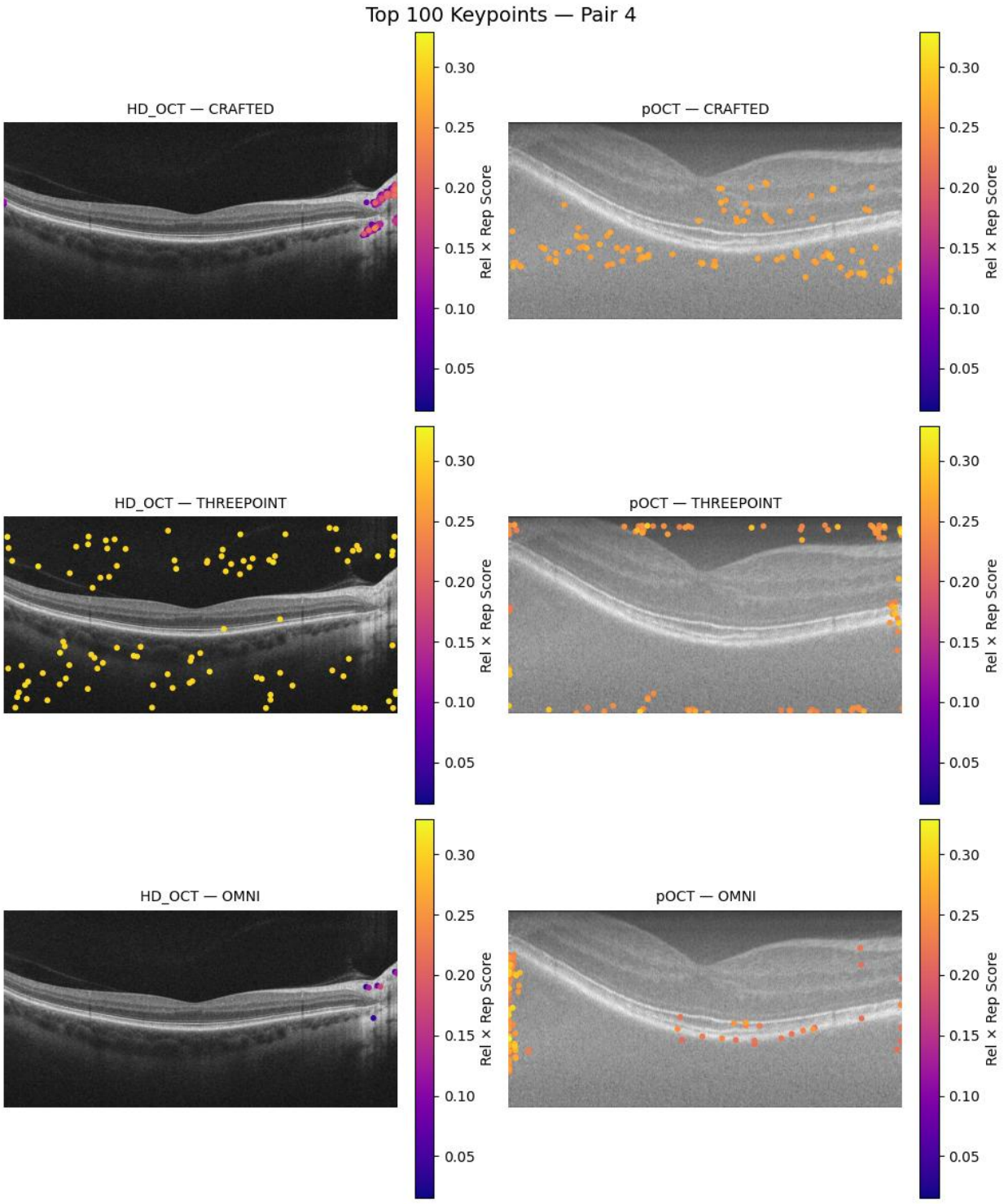

Figure 18
Top 100 keypoints visualized for Pair 4 across all models and modalities. The CRAFTED model produces anatomically plausible keypoints along the RPE in HD-OCT, but its pOCT output is spatially diffuse and lacks correspondence. The THREEPOINT model shows a consistent distribution across modalities, but fails to target retinal structures, reducing its geometric value. The OMNI model produces moderately reliable keypoints in pOCT, but yields few or no meaningful keypoints in HD-OCT, eliminating the possibility of defining an accurate transformation. Overall, these results underscore the difficulty of establishing cross-modality correspondence even when intra-modality repeatability and reliability appear adequate.

Chapter 5

Registration

Following the extensive evaluation of keypoints through both quantitative and qualitative metrics for all three models, the next logical step was to assess the utility of these keypoints and their corresponding descriptors in the context of the registration task that is the main purpose of this thesis. While a thorough analysis of repeatability and reliability scores and ability to describe anatomically relevant areas is vital in determining the optimal approach, the ultimate test of a keypoint detection and description system lies in its capacity to drive robust and accurate image alignment. This chapter introduces and implements a registration pipeline that leverages the previously extracted keypoints and descriptors to estimate inter-device retinal image alignment. Firstly, a traditional approach is implemented that uses the existing models in a straightforward manner and later a fusion of the models to take advantage of their independent strengths is explored to determine if it can produce better results.

5.1 Descriptor Based Registration Using Three Different Models

As already applied in earlier iterations of R2D2 in image registration pipelines, the process is structured around three main pillars, Euclidean distance, geometric model estimation via RANSAC and image alignment through affine transformation. [56]

Euclidean distance refers to a similarity matrix valuable in machine learning and computer vision. The Euclidean distance d of two data cases (x_1, x_2) is defined as the square root of the sum of the squared differences.

$$d(x_i, y_i) = \sqrt{\sum |x_i - y_i|^2}$$

In high dimensional descriptor matching tasks, as is the one tackled here, Euclidean distance is used as a quantitative measure of similarity between the keypoints that are described by the descriptor vectors. A lower distance implies greater similarity between descriptors, which in theory can be translated, as a criterion to drive matches between images. In our context, Euclidean distance is used to compare the similarities between our 128-dimensional descriptors. Mutually closest pairs, which means that if descriptor A in the first image has closest to it in terms of Euclidean distance descriptor B of the second image, then descriptor B must also have descriptor A as its closest, will reveal a small set of correspondences that will serve as drivers of transformation estimation in the registration pipeline. [64]

Random Sample Consensus, which is also abbreviated as RANSAC, is a robust estimation method for models that require fitting in the presence of outliers. It creates several models in succession, each containing a random number of samples each time. Each model is fitted into each subset, and then the number of points that agree with the model based on a predefined threshold value. The model with the highest number of inliers is considered the best. It can vital in estimating the affine transformation needed to guide image alignment as many corresponding keypoints may be output by our Euclidean distance metric and RANSAC can reliably filter out any unreliable matches.[55, 65]

Hence, each image pair had its descriptors loaded that corresponded to the pixel location of the checkpoints. Descriptor matching was performed using their Euclidean distance between the HD-OCT and pOCT images. A mutual nearest neighbor strategy was employed, as described previously, which helped eliminate ambiguous matches and reduce the possibilities of erroneous correspondences. Following this, the corresponding keypoint coordinates of the descriptor matches were extracted and through the application of RANSAC an optimal affine transformation was calculated. If less than three inliers were found, the affine transformation calculation was not possible as described earlier. This registration pipeline took place independently for each model across all 84 intermodal pairs. Also, the number of mutual matches, the mutual match ratio - the number of mutual matches divided by the number of total descriptors – and the average and standard deviation of the Euclidean distance for each model were calculated. A maximum of the 50 matches was applied to avoid increased noise.

Quantitative analysis of the descriptors match revealed substantial differences in behavior among C, 3P and O. On average, the number of mutual matches per image pair was highest for 3P (50), followed by C (6.3), and lowest for O (1.095). This suggests that while 3P produces a larger number of descriptor correspondences, O often fails to generate enough confident matches to guide reliable image alignment. However, when considering the mutual match ratio—defined as the proportion of descriptors from the smaller descriptor set that participate in mutual nearest neighbor matches—C exhibited the highest value (0.509), whereas 3P had the lowest (0.01). This indicates that although 3P produces more matches in absolute terms, only a small fraction of its descriptors is successfully matched, potentially due to an overabundance of non-discriminative or redundant features. In contrast, C's high match ratio reflects a more efficient and targeted descriptor set. Lastly, the average Euclidean distance between matched descriptors further illustrates this divergence: 3P had the lowest mean distance (0.154), suggesting more similar or tightly clustered descriptors, while C had the highest (0.436), potentially implying poorer descriptor precision or noisier matches. Taken together, these metrics suggest that while 3P excels in producing numerous close descriptor matches, C maintains a higher match efficiency, and O underperforms in both aspects.

Following the matching process, affine transformations were estimated using RANSAC for each image pair based on the mutually matched descriptors. After this process, if inlier matches were less than 3 the models could not produce a transformation matrix and thus in these instances the model failed. Specifically, C failed in 46.4% of cases, O in 92.8% of cases while 3P achieved a 100% success rate. Arbitrarily, pOCT were assigned the role of the moving image and HD-OCT that of the fixed image. The resulting affine matrices were applied to produce a warped moving image. Figures 19 contains representative examples of the registration results, illustrating how the models would transform the images to achieve registration.

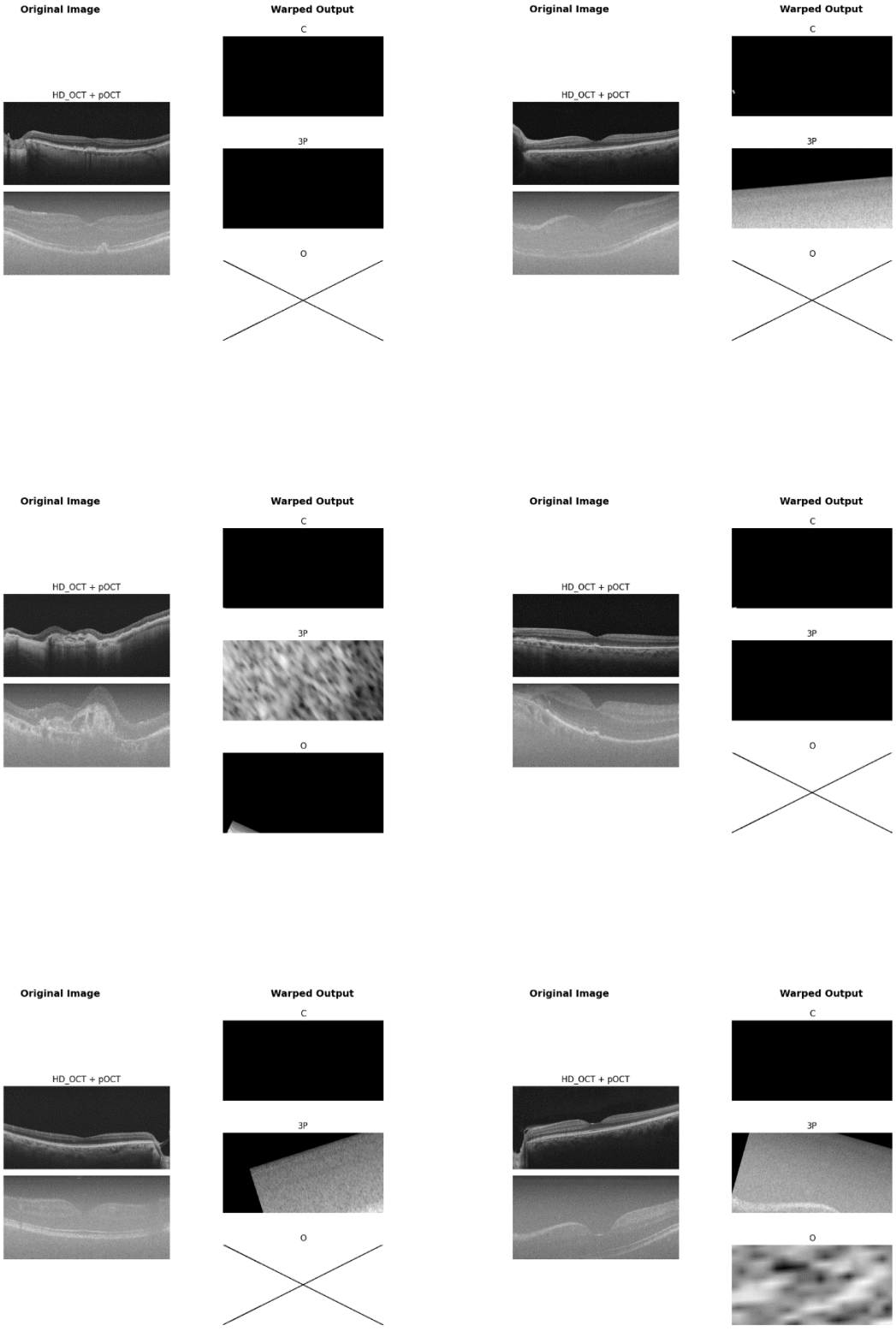

Figure 19
Representative registration results from all three models across six image pairs. In each case, the computed affine transformations fail to achieve meaningful alignment, highlighting the models' inability to guide successful image registration.

A visual qualitative inspection of the applied transformations revealed a consistent and substantial failure to achieve anatomical alignment. In all the instances, the warped images bore no resemblance to the fixed images, with significant distortions and extremely aggressive transformations applied. This discrepancy suggests that the rationale before applying the matching of keypoints may be flawed in our instance. The O model displayed the worst performance by not being able to produce a transformation in almost all cases, due to limited number of keypoints, matches or inliers. In the few instances that the model produced a transformation, it was extreme and resulted in tremendous scaling changes, which in no way correspond to true alignment with the original image. On the other hand, the 3P managed to produce a transformation matrix in all cases, but did not perform better in terms of alignment. Once more, transformations were aggressive in scaling or overly skewed, without the ability to align corresponding anatomical landmarks. Finally, the C model, despite yielding a higher match ratio, was also unable to produce meaningful transformation, following the same shortcomings as the previous two.

These observations solicited a careful inspection of the derived transformation matrices to examine how the models did not manage the required task. During this examination, the initial hypothesis that all models applied aggressive transformations was further validated. Specifically, 3P and O model provided translations that often exceeded 5000 pixels, near 5 times the size of the image, which in no way can correspond to real anatomical alignment. These extreme values also explain why many warped images were completely empty after transformation, as the original image was shifted outside the original frame. The C model, while more conservative, also exceeded 500 pixels in translation, significantly overshooting any meaningful transformations. Scaling was also aggressive, reaching values of 20x, and inevitably producing unreliable results. Finally, rotation estimates were similarly erratic across all models. While the median rotation for each model were modest, the distribution was marked by extremes, numerous cases exceeding 90 degrees and several flipping the image completely upside down. The combination of these extreme metrics can in no way reveal meaningful transformations.

The complete failure of all three models in such a similar manner suggests that the root of the issue lies on the matching process rather than model performance. This conclusion is derived based on the significant differences in model behavior both quantitatively and qualitatively, as reported previously, which comes in stark contrast with the complete collapse of all models during the registration process in almost identical ways. This can be further reviewed by examining the matches that were procured after applying the rationale of mutually exclusive pairs guided by Euclidean distance.

Summary of Affine Transformation Metrics by Model						
Model	**Mean Translation (px)**	**Max Translation (px)**	**Mean Scale**	**Max Scale**	**Mean Rotation (°)**	**Rotation Std Dev (°)**
Crafted	649.14	1769.12	0.25	2.23	8.95	98.23
Threepoint	970.97	7720.44	1.34	13.31	−9.30	91.08
Omni	1259.47	5061.63	4.36	22.51	−35.96	65.84

Table 3

Summary of key geometric characteristics derived from the affine transformation matrices computed by each model

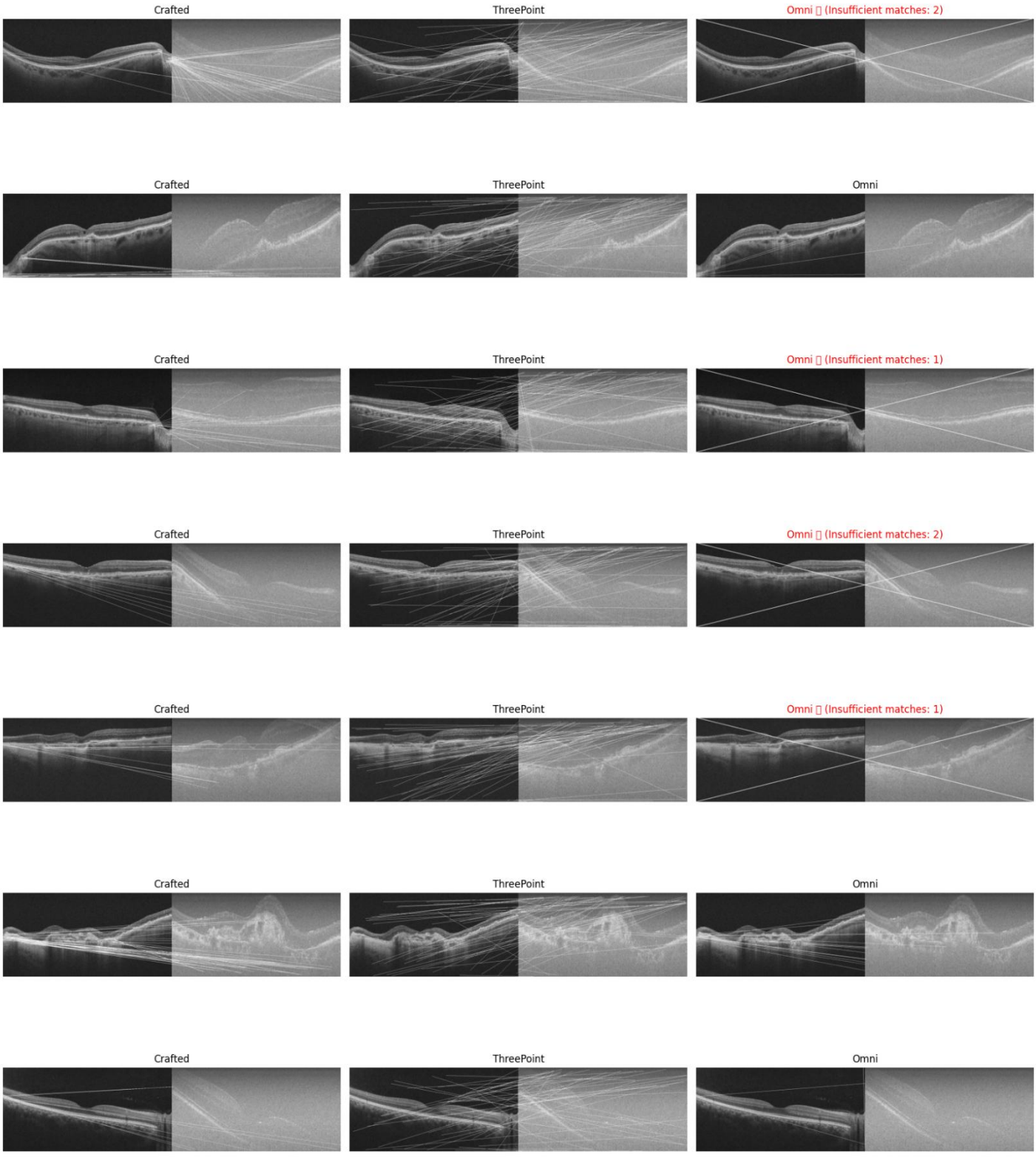

Figure 20
Qualitative visualization of keypoint matching results across three models (Crafted, ThreePoint, and Omni) for seven representative image pairs. Each row corresponds to a different pair, with models compared left to right. Lines indicate matched descriptors between fixed and moving images

Upon careful inspection of the matched keypoints the descriptor Euclidean distance produced, it is easy to discern the root cause of the misalignment produced during the traditional approach of the registration process. Figure 20 shows some representative pairs across all models. The matched keypoints rarely correspond to corresponding anatomical areas. Instead, there is there is a recurring pattern of spatial inconsistency. Also, it is often observed that keypoints that describe a certain area of a retinal layer in one modality, i.e. the nasal side of the RPE, scatter across the entirety of the RPE or even entirely irrelevant anatomies in the other modality and vice versa, indicating a flaw in the rationale of keypoint matching as this pattern of matching cannot possibly produce meaningful results. The matching logic of the mutual nearest neighbors in Euclidean descriptor space assumes that similar descriptors represent similar anatomy. However, it is evident in our case that similar descriptors frequently correspond to structurally unrelated or distant regions, hence compromising alignment robustness.

The Euclidean distance approach assumes that descriptor similarity correlates with anatomical correspondence across images. As is clearly demonstrated, this assumption does not bear truth in multimodal image registration as our images exhibit varying contrast levels and significant noise artifacts. One of the key limitations arises in this setting arises from the sensitivity of Euclidean distance to local intensity variations. A subset of our images exhibits speckle noise – a granular interference patter that reduces image quality – which can cause descriptors to become distorted and less distinctive. Descriptors can become biased by local noise and reflect that during the comparison process, where noise distortions can produce matches that do not correspond to true anatomical similarity but rather similar intensity between images. Hence, erroneous matches are created which lead to the disappointing results presented earlier. These observations point to a deeper limitation in similarity metrics and traditional approaches in medical image registration where noise is present and call for an alternative approach to better utilize the advantages of trained models. [66-69]

5.2 Fusion Approach to Registration

As the results of the traditional application of the R2D2 models thus far returned underwhelming results, an alternative approach is required to proceed. The proposed approach aims to use all three fine-tuned models and leverage its advantages to achieve the optimal result.

Initially, a pressing matter that was discovered during the earlier iterations was the assignment of keypoints to areas with little to no anatomical relevance to the retinal contour. As it was earlier observed, the 3P model exhibited very stark contrasts in reliability scores when transitioning form the hyaloid to the ILM and also when transitioning from areas posterior to the choroid to the choroid or RPE. This difference in reliability values is more easily recognizable in HD-OCT reliability maps as also shown in Fig.16 and 17. These two interfacces motivated the use of a column-wise gradient based detection algorithm to localize the ILM and the RPE and choroid limits using only the model's reliability output. Specifically, the methodology involved processing the HD-OCT images from the dataset and the reliability maps were produced. Each column of said map was scanned from top to bottom to locate the first band of 10 or more consecutive pixels with reliability scores below 0.05. Those values were determined after observation of the map output and after other values were also tried and disregarded. The zone found would represent the ILM.

A similar approach was used, scanning from the bottom-up to locate the choroid or the RPE. Also the restriction of the RPE or choroid layer to be located at least 20 pixels below the corresponding ILM was applied. To ensure continuity across columns and avoid anatomical discontinuity due to the potenital inability of the approach to locate the wanted band, missing ILM and RPE/Choroid values were interpolated using the closest neighboring valid values. Once the two boundaries were finalized, a binary gate mask was constructed by marking as valid pixels all pixels between the two boundaries. The results, while not perfect, manage to contain the majority of the later derived keypoints within anatomically relevan regions.

Attempts to apply the same logic in pOCT images failed the qualitative control. The gradient is not that evident in any of the three models' repeatability or reliability maps. However, an approach was needed to ensure that keypoints far away from the retinal contour were excluded. To adress this challenge, an alternative masking strategy was employed. While the retinal boundaries were not as clearly demarcated, the generalized pattern remained the same; areas relevant to retinal contour were assigned lower reliability valued compared to their hyaloid or posterior to the choroid areas. This observation led to a patch-based masking strategy that favors areas with lower assigned reliability scores. In this approach, the pOCT reliability map is divided vertically into non-overlapping patches of 80 pixels in height. For every patch, we count the number of pixels whose reliability score falls below a certain threshold and assign a percentage that needs to be under that threshold to qualify as a relevant patch. After careful inspection of reliability maps and the trial of many alternative values, the optimal for our scenario were the threshold of value 0.4 and at least 50 of the 80 pixels to be assigned a lesser than threshold value. Overall, the 3P model contributed to this approach by setting such boundaries which would ideally constrain the allocation of keypoints to anatomically rich in information region.

Our previous qualitative evaluation of repeatability and reliability maps revealed great performance by the C model on all repeatability maps and an enhanced prowess of the repeatability maps of the O model towards pOCT images. To leverage both models' strenghts, fused repeatability maps were computed for the two models. The repeatability maps were extracted and added element-wise to form a fused maps. Scores were added to produce the final score. The multiplication of the scores was avoided as O frequently assigned 0 values and hence would completely alleviate the stellar scores the C model assigned in some instances.

The extracted keypoints were filtered based on the 3P model's masks. Each fused repeatabiliy map was then masked by its corresponding anatomical gate following the following pattern.

$$R_{filtered}(x, y) = R_{fused}(x, y) \cdot M(x, y)$$

This ensured that only points within the previously defined masks were eligible to be considered keypoints.

From each filtered fuse map a similar logic was applied for the extraction of keypoints as before. Specifically, 400 top scoring keypoints were extracted using non-maximum suppression over a 3x3 window. Only local maxima were candidates. Keypoints were thus extracted independently for both HD-OCT and pOCT images. To avoid the Euclidean distance shortcomings

described earlier, only the coordinates and associated scores of the keypoints were saved and not their 128-dimensional descriptors.

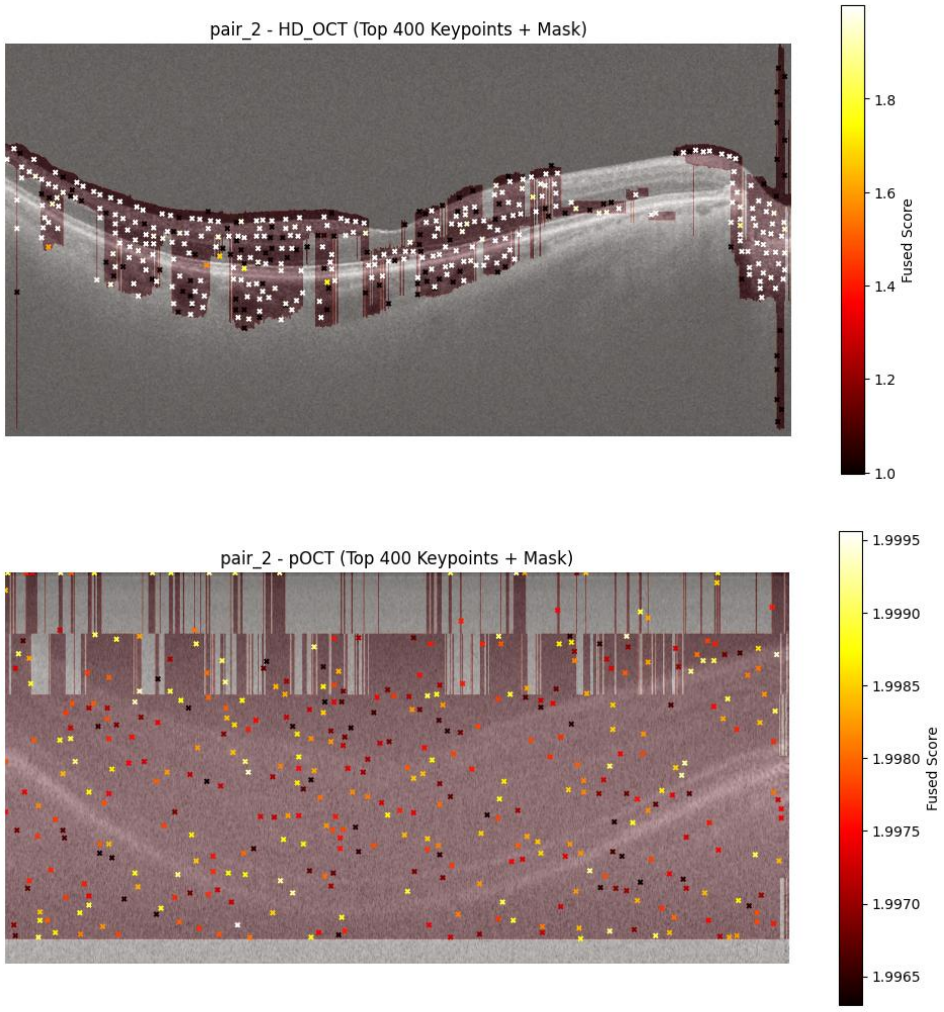

Figure 21
Representative example of the fusion model approach. Keypoints are overlaid on the images and color coded based on their derived fused score. The red hue represents the mask applied to each image. While not perfectly demarcating the retinal structure boundaries, it manages to significantly reduce the amount of potential keypoints outside the retinal contour

To implement keypoint matching a more straightforward approach of spatial nearest-neighbor was applied. For each image in one modality the nearest keypoint in its counterpart was found using a k-d tree structure. Similarly to the traditional approach, RANSAC was employed to identify inliers and compute the necessary affine transformation to warp the pOCT images. They were chosen for consistency and ease of comparison with the previous model. In this instance, all image pairs had at least 3 inliers which resulted in no failed registration attempts. [70, 71]

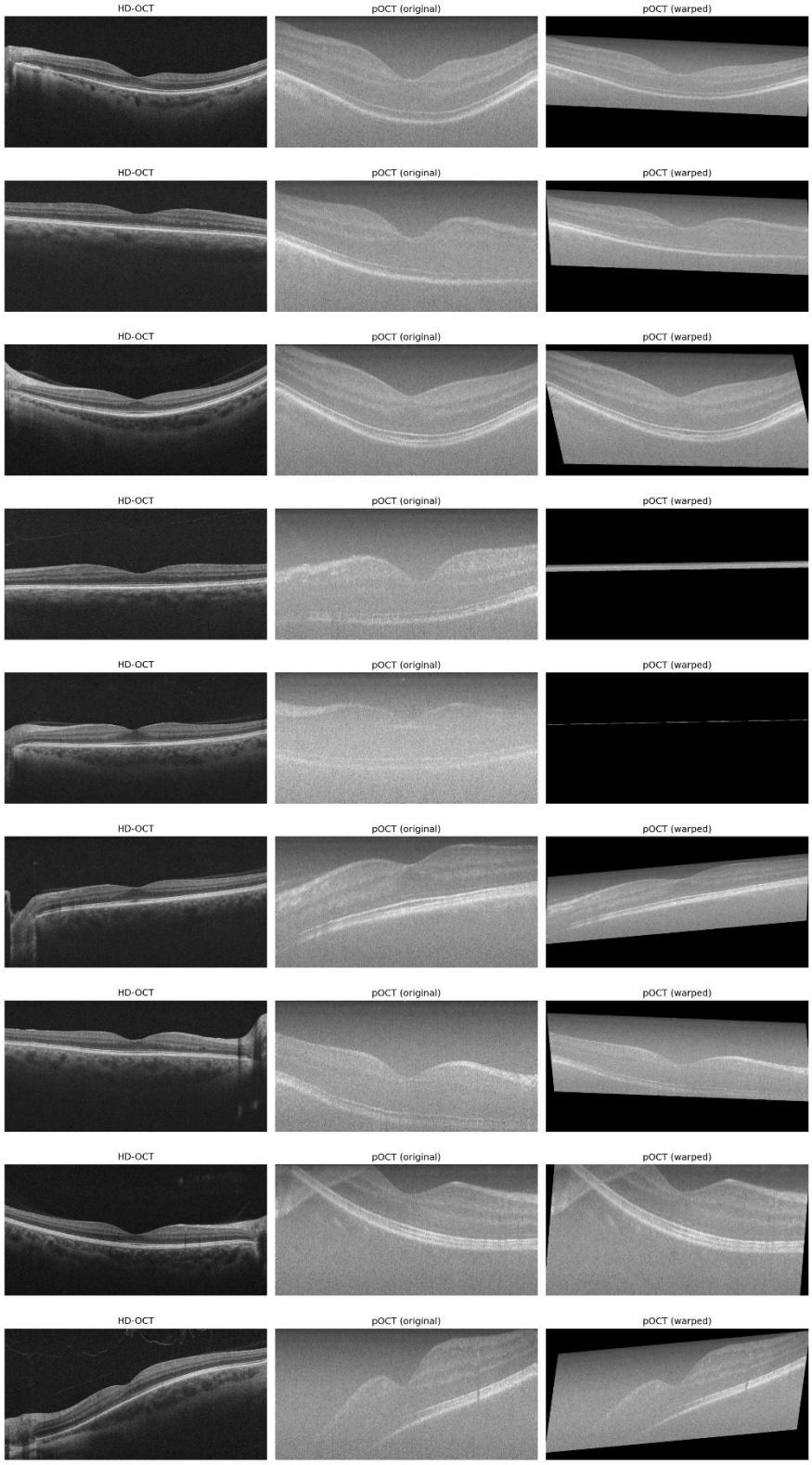

Figure 22
Representative examples of affine alignment using the fusion approach. While not perfect they outperform the traditional approach and, in some cases, reach acceptable results. However, exaggerated transformations that distort the warped image can also be found here.

5.2 Fusion Approach to Registration

Upon initial inspection, the fusion approach significantly outperforms the traditional descriptor-Euclidean distance approach. Images are always in frame and in some cases, results are close to excellent. Careful inspection of the transformation matrices backs up this initial assessment. The translations are limited to a maximum of 432 pixels compared to the 7000 of the worst performing O model and also scaling is significantly more conservative with a mean value of 0.83 compared to the very small value of 0.25 of the crafted model and the aggressive 4.36 of the omni model. Additionally, rotation had an extremely tight distribution that prevented significant misalignment of the warped image. Overall, the fused model is vastly more conservative and controlled across all transformation parameters, avoiding erratic and aggressive transformation. This conservative approach is the reason that some pairings produce an acceptable result that can be used in our pipeline of eventual super resolution. Such examples can be seen in a checkerboard form in figure 23.

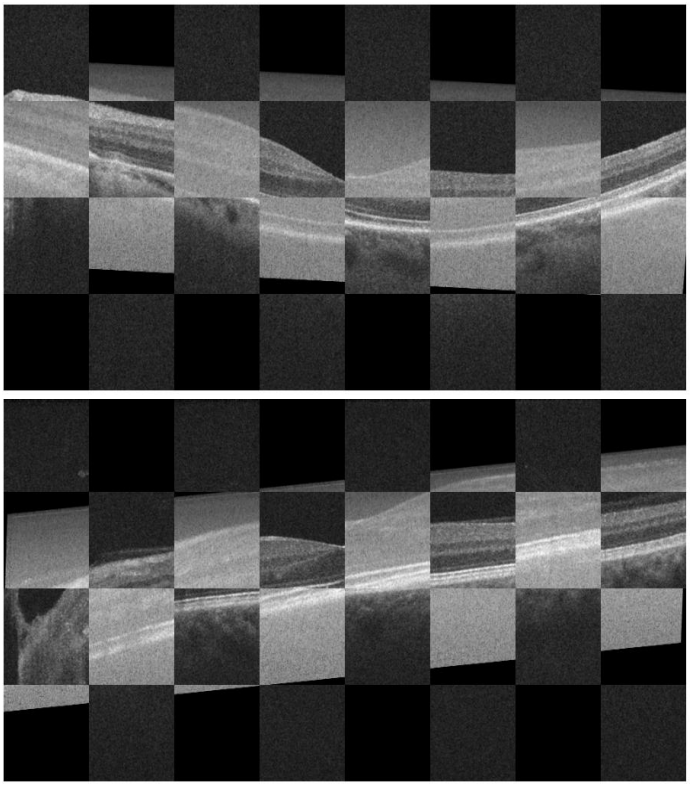

Figure 23

The two most successful example of the fused model application, showing both the HD_OCT and the transformed pOCT image in a checkerboard composite. While not perfect they manage to closely align the main retinal structures with slight vertical misalignment of the layers. It is a promising beginning that possibly suggest that future minor alterations may produce a more robust and reproducible result.

Chapter 6

Discussion and Future Directions

This thesis set out to evaluate the feasibility of registering inter device OCT images through a deep learning keypoint detection and description strategy. By recruiting the R2D2 framework the possibility of deriving transformation that aligned images across different imaging modalities was explored. The results yielded suggest that even in the challenging setting of multimodal OCT registration, a learned detector-descriptor model can provide meaningful correspondences, especially after careful evaluation of repeatability and reliability maps.

Multimodal OCT registration is inherently challenging due to a variety of factors. Firstly, OCT images are characterized by texture-based ambiguity and presence of speckle noise. To make matters worse, texture and noise can vary across devices, thus rendering correspondences even more difficult. Additionally, OCT images lack the crystal-clear anatomical landmarks that other ophthalmologic examinations provide such as fundus photography or corneal pachymetry maps. Given these limitations, previously adopted approaches in similar tasks and field, failed to produce a reliable result in our case. A potential exploration of a different similarity metric, other than Euclidean distance, may be the next step in optimizing the utilization of derived descriptors.

Extremely promising was the valuable results derived by considering two images derived from different modalities as essentially equally and treating them as such in the training pipeline. Introducing expert annotated correspondences between the HD-OCT and pOCT images through this process allowed for the development of two extra models, which provided invaluable information in the development of the fused model. Additionally, the minimum input that experts need to make – only three corresponding points between the images- is enough to generate multiple new pairs for training, essentially providing the foundation for a semi-supervised approach, as the correspondence between the original and warped image is calculated immediately after the initial application of the transformation.

Among all the models tested, a fusion of the C, 3P and O yielded the most usable results. This further highlights the importance of the rationale of considering two images of different modalities as equal after initial expert driven rough alignment. Each model on its own however failed to produce any meaningful results. Quantitatively, 3P achieved the greatest number of keypoints, while C and O struggled to produce keypoints high in reliability values. The distribution of the C keypoints when it came to repeatability scores greatly captured the retinal contour and was a deciding factor in the level of “success “the fused model achieved.

Despite promising outcomes, several limitations remain. As this was a study in the feasibility of multimodal image registration, the generalizability of the proposed pipeline was not evaluated in this thesis. This work acts as a proof of plausibility rather than a general-purpose solution. The next logical step would be to apply the rationale of the fusion model to an external dataset of paired OCT scans of the same or different modalities to test for potential overfitting in our approach.

Another limitation is that our current pipeline does not integrate anatomical annotations such as retinal layer boundaries. This happened intentionally to assess the plausibility of only a

deep learning registration approach. The incorporation of segmentation masks can greatly improve the outcome of our framework, as one of the most limiting steps in our pipeline was to determine a sufficient method to exclude keypoints that returned no useful anatomical transformation. However, such an approach introduces the need for expert annotations that require a significant amount of time and effort. This contradicts in principle the philosophy of this thesis and also is unnecessarily resource intensive, as we have already established that passable alignment can happen by simply annotating only three corresponding keypoints. This obstacle can be tackled by also developing a deep learning layer segmentation model that will be able to automatically annotate retinal layers across different modalities and thus keep the process annotation-free. Such an effort is a logical next step and annotation of layers in the existing dataset has already begun towards this end.

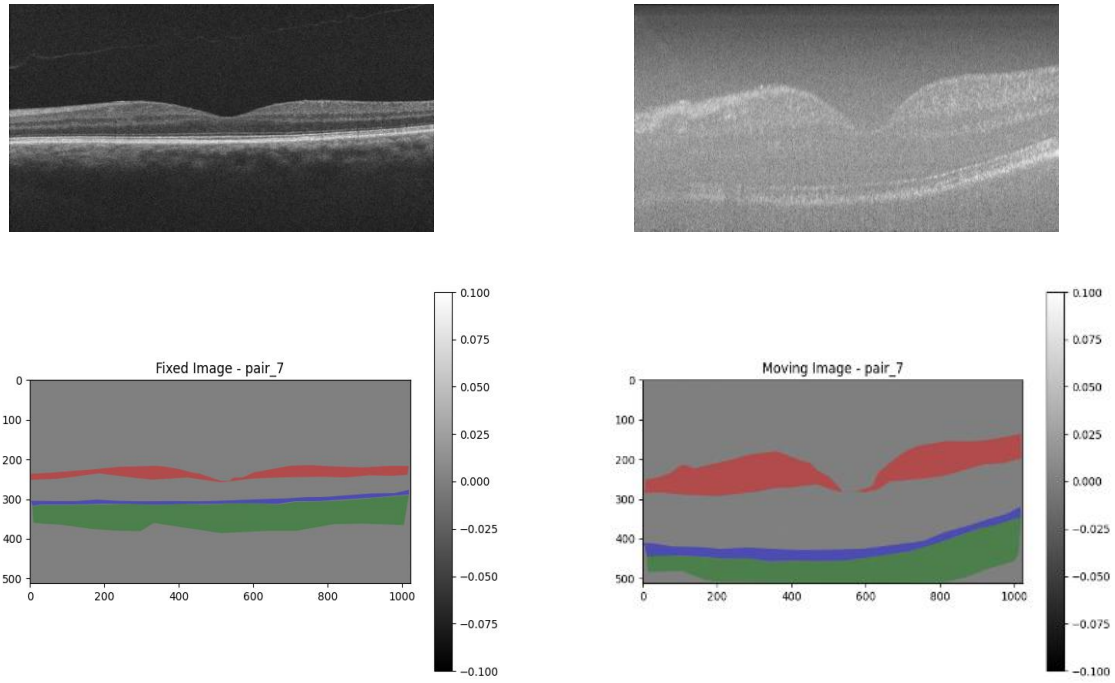

Figure 24
Example of already existing retinal layer annotations in an effort to develop a layer segmentation framework to facilitate accurate masking

Finally, it is important to mention that this thesis in no way is ready to act as a clinical registration tool. Its contribution lies on the demonstration of the feasibility of extracting keypoints and descriptors both repeatable and reliable that have the potential to guide successful OCT inter device image registration, should certain conditions be met. These findings allow for future exploration of a different approach in matching keypoints or descriptors, alternative ways of introducing masking or approaching the issue in a totally new manner as it is already established that fine-tuned R2D2 is a valuable ally in this endeavor. Such a breakthrough would bring the possibility of automated multimodal OCT image registration closer to the clinical setting and help bridge the gap between deep learning, image analysis and clinical ophthalmic practice.

Bibliography

[1] C. Zhou, S. Li, L. Ye, C. Chen, S. Liu, H. Yang*, et al.*, "Visual impairment and blindness caused by retinal diseases: A nationwide register-based study," *J Glob Health*, vol. 13, no. pp. 04126,2023 doi: 10.7189/jogh.13.04126.

[2] M. Fleckenstein, S. Schmitz-Valckenberg, and U. Chakravarthy, "Age-Related Macular Degeneration: A Review," *Jama*, vol. 331, no. 2, pp. 147-157,2024 doi: 10.1001/jama.2023.26074.

[3] F. J. Rodríguez, G. Staurenghi, and R. Gale, "The role of OCT-A in retinal disease management," *Graefes Arch Clin Exp Ophthalmol*, vol. 256, no. 11, pp. 2019-2026,2018 doi: 10.1007/s00417-018-4109-3.

[4] R. Chopra, S. K. Wagner, and P. A. Keane, "Optical coherence tomography in the 2020s-outside the eye clinic," vol. 35, no. 1, pp. 236-243,2021 doi: 10.1038/s41433-020-01263-6.

[5] G. Song and E. T. Jelly, "A review of low-cost and portable optical coherence tomography," vol. 3, no. 3, pp.,2021 doi: 10.1088/2516-1091/abfeb7.

[6] S. Umirzakova, S. Mardieva, and S. Muksimova, "Enhancing the Super-Resolution of Medical Images: Introducing the Deep Residual Feature Distillation Channel Attention Network for Optimized Performance and Efficiency," vol. 10, no. 11, pp.,2023 doi: 10.3390/bioengineering10111332.

[7] K. Yamashita and K. Markov, *Medical Image Enhancement Using Super Resolution Methods*: Computational Science – ICCS 2020. 2020 May 25;12141:496-508. doi: 10.1007/978-3-030-50426-7_37.

[8] K. A. Thakoor, A. Carter, G. Song, A. Wax, O. Moussa, R. W. S. Chen*, et al.*, "Enhancing Portable OCT Image Quality via GANs for AI-Based Eye Disease Detection," Cham, 2022, pp. 155-167.

[9] P. E. Ludwig, R. Jessu, and C. N. Czyz, "Physiology, Eye," in *StatPearls*, vol., Ed.^Eds., ed., Treasure Island (FL): StatPearls Publishing, 2025.

[10] H. Kolb, "Simple Anatomy of the Retina," in *Webvision: The Organization of the Retina and Visual System*, vol., H. Kolb, E. Fernandez, B. Jones, and R. Nelson, Ed.^Eds., ed., Salt Lake City (UT): University of Utah Health Sciences Center, 1995.

[11] B. D. Kels, A. Grzybowski, and J. M. Grant-Kels, "Human ocular anatomy," *Clin Dermatol*, vol. 33, no. 2, pp. 140-6,2015 doi: 10.1016/j.clindermatol.2014.10.006.

[12] M. Fleckenstein, T. D. L. Keenan, R. H. Guymer, U. Chakravarthy, S. Schmitz-Valckenberg, C. C. Klaver*, et al.*, "Age-related macular degeneration," *Nature Reviews Disease Primers*, vol. 7, no. 1, pp. 31,2021 doi: 10.1038/s41572-021-00265-2.

[13] S. Aumann, S. Donner, J. Fischer, and F. Müller, "Optical Coherence Tomography (OCT): Principle and Technical Realization," in *High Resolution Imaging in Microscopy and Ophthalmology: New Frontiers in Biomedical Optics*, vol., J. F. Bille, Ed.^Eds., ed., Cham (CH): Springer, 2019.

[14] B. E. Bouma, J. F. de Boer, D. Huang, I. K. Jang, T. Yonetsu, C. L. Leggett*, et al.*, "Optical coherence tomography," *Nat Rev Methods Primers*, vol. 2, no. pp.,2022 doi: 10.1038/s43586-022-00162-2.

[15] K. Irsch, "Optical Principles of OCT," in *Albert and Jakobiec's Principles and Practice of Ophthalmology*, vol., D. Albert, J. Miller, D. Azar, and L. H. Young, Ed.^Eds., ed., Cham: Springer International Publishing, 2020.

[16] J. S. Schuman, "Spectral domain optical coherence tomography for glaucoma (an AOS thesis)," *Trans Am Ophthalmol Soc*, vol. 106, no. pp. 426-58,2008 doi:

[17] F. Xia and R. Hua, "The Latest Updates in Swept-Source Optical Coherence Tomography Angiography," vol. 14, no. 1, pp.,2023 doi: 10.3390/diagnostics14010047.

[18] M. Bhende, S. Shetty, M. K. Parthasarathy, and S. Ramya, "Optical coherence tomography: A guide to interpretation of common macular diseases," *Indian J Ophthalmol*, vol. 66, no. 1, pp. 20-35,2018 doi: 10.4103/ijo.IJO_902_17.

[19] M. Chen, N. J. Tustison, R. Jena, and O. Colliot, "Image Registration: Fundamentals and Recent Advances Based on Deep Learning," in *Machine Learning for Brain Disorders [Internet]*, vol., O. Colliot, Ed.^Eds., ed., United States: Humana, 2023.

[20] J. Liu, G. Singh, S. Al'Aref, B. Lee, O. Oleru, J. K. Min*, et al.*, "Image Registration in Medical Robotics and Intelligent Systems: Fundamentals and Applications," *Advanced Intelligent Systems*, vol. 1, no. 6, pp. 1900048,2019 doi: https://doi.org/10.1002/aisy.201900048.

[21] V. B. Sivaraman, M. Imran, Q. Wei, P. Muralidharan, M. R. Tamplin, I. M. Grumbach*, et al.*, "RetinaRegNet: A zero-shot approach for retinal image registration," *Computers in Biology and Medicine*, vol. 186, no. pp. 109645,2025 doi: https://doi.org/10.1016/j.compbiomed.2024.109645.

[22] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Deep learning in medical image registration: a review," *Phys Med Biol*, vol. 65, no. 20, pp. 20tr01,2020 doi: 10.1088/1361-6560/ab843e.

[23] P. Arora, R. Mehta, and R. Ahuja, "An adaptive medical image registration using hybridization of teaching learning-based optimization with affine and speeded up robust features with projective transformation," *Cluster Computing*, vol. 27, no. 1, pp. 607-627,2024 doi: 10.1007/s10586-023-03974-3.

[24] C. Li, J. Sun, X. Zhang, L. Zhang, X. Sun, and L. Wang, "An seamless stitching method for large field equivalent center projection image based on rotating camera," *Sci Rep*, vol. 14, no. 1, pp. 29170,2024 doi: 10.1038/s41598-024-80295-4.

[25] Y. Rong, M. Rosu-Bubulac, S. H. Benedict, Y. Cui, R. Ruo, T. Connell*, et al.*, "Rigid and Deformable Image Registration for Radiation Therapy: A Self-Study Evaluation Guide for NRG Oncology Clinical Trial Participation," *Pract Radiat Oncol*, vol. 11, no. 4, pp. 282-298,2021 doi: 10.1016/j.prro.2021.02.007.

[26] M. Abdel-Basset, A. E. Fakhry, I. El-Henawy, T. Qiu, and A. K. Sangaiah, "Feature and Intensity Based Medical Image Registration Using Particle Swarm Optimization," *J Med Syst*, vol. 41, no. 12, pp. 197,2017 doi: 10.1007/s10916-017-0846-9.

[27] A. Myronenko and X. Song, "Intensity-based image registration by minimizing residual complexity," *IEEE Trans Med Imaging*, vol. 29, no. 11, pp. 1882-91,2010 doi: 10.1109/tmi.2010.2053043.

[28] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, "LF-Net: Learning local features from images," *Advances in neural information processing systems*, vol. 31, no. pp.,2018 doi:

[29] A. Okorie and S. Makrogiannis, "Region-based image registration for remote sensing imagery," *Computer Vision and Image Understanding*, vol. 189, no. pp. 102825,2019 doi: https://doi.org/10.1016/j.cviu.2019.102825.

[30] E. Rosten, R. Porter, and T. Drummond, "Faster and better: a machine learning approach to corner detection," *IEEE Trans Pattern Anal Mach Intell*, vol. 32, no. 1, pp. 105-19,2010 doi: 10.1109/tpami.2008.275.

[31] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110,2004 doi: 10.1023/B:VISI.0000029664.99615.94.

[32] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in *Computer Vision – ECCV 2010*, Berlin, Heidelberg, 2010, pp. 778-792.

[33] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*, 2011, pp. 2564-2571.

[34] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A Learning Framework for Deformable Medical Image Registration," *IEEE Trans Med Imaging*, vol. no. pp.,2019 doi: 10.1109/tmi.2019.2897538.

[35] 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. 337-33712.

[36] P. E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning Feature Matching With Graph Neural Networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4937-4946.

[37] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, *LightGlue: Local Feature Matching at Light Speed*, 2023.

[38] C. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal Correspondence Network," vol. no. pp.,2016 doi: 10.48550/arXiv.1606.03558.

[39] K. Yi, E. Trulls, V. Lepetit, and P. Fua, *LIFT: Learned Invariant Feature Transform* vol. 9910, 2016.

[40] J. Revaud, P. Weinzaepfel, C. De Souza, N. Pion, G. Csurka, Y. Cabon*, et al.*, *R2D2: Reliable and Repeatable Detectors and Descriptors for Joint Sparse Keypoint Detection and Local Feature Extraction*, 2019.

[41] Y. Tian, B. Fan, and F. Wu, *L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space*, 2017.

[42] R. Bhuiyan, J. Abdullah, N. Hashim, and F. Al Farid, "Deep Dilated Convolutional Neural Network for Crowd Density Image Classification with Dataset Augmentation for Hajj Pilgrimage," vol. 22, no. 14, pp.,2022 doi: 10.3390/s22145102.

[43] E. Zhang and Y. Zhang, "Average Precision," in *Encyclopedia of Database Systems*, vol., L. Liu and M. T. ÖZsu, Ed.^Eds., ed., Boston, MA: Springer US, 2009.

[44] A. Husham Al-Badri, N. Azman Ismail, K. Al-Dulaimi, G. Ahmed Salman, and M. Sah Hj Salam, "Adaptive Non-Maximum Suppression for improving performance of Rumex detection," *Expert Systems with Applications*, vol. 219, no. pp. 119634,2023 doi: https://doi.org/10.1016/j.eswa.2023.119634.

[45] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg*, et al.*, *Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions*, 2018.

[46] F. Lu, D. Zhou, H. Chen, S. Liu, X. Ling, L. Zhu*, et al.*, "S2P-Matching: Self-Supervised Patch-Based Matching Using Transformer for Capsule Endoscopic Images Stitching," *IEEE Transactions on Biomedical Engineering*, vol. 72, no. 2, pp. 540-551,2025 doi: 10.1109/TBME.2024.3462502.

[47] J. Wang, H. Li, D. Hu, R. Xu, X. Yao, Y. K. Tao*, et al.*, "Retinal IPA: Iterative KeyPoints Alignment for Multimodal Retinal Imaging," in *Medical Optical Imaging and Virtual Microscopy Image Analysis*, Cham, 2025, pp. 119-129.

[48] L. Zeyuan, Z. Zirui, C. Wenguang, W. Yi, C. Huaiyu, and C. Xiaodong, "A novel learning-based keypoint matching framework for toric intraocular lens navigation during cataract surgery," in *Proc.SPIE*, 2024, p. 1323907.

[49] W. Zhao, X. Xu, J. Xie, L. Cheng, and Z. Zhang, "Detection Method of Eye Rotation Angle in Cataract Surgery Based on Depth Feature Matching," *Journal of Computer-Aided Design & Computer Graphics*, vol. 36, no. 9, pp. 1407-1417,2024 doi: 10.3724/SP.J.1089.2024.19997.

[50] D. Rivas-Villar, Á. S. Hervella, J. Rouco, and J. Novo, "ConKeD: multiview contrastive descriptor learning for keypoint-based retinal image registration," *Medical & Biological Engineering & Computing*, vol. 62, no. 12, pp. 3721-3736,2024 doi: 10.1007/s11517-024-03160-6.

[51] M. Sommersperger, P. Matten, T. Wang, S. Dehghani, J. Nienhaus, H. Roodaki*, et al.*, "Context-aware real-time semantic view expansion of intraoperative 4D OCT," *IEEE Transactions on Medical Imaging*, vol. no. pp. 1-1,2025 doi: 10.1109/TMI.2025.3528742.

[52] Y. Hu, M. Gong, Z. Qiu, J. Liu, H. Shen, M. Yuan*, et al.*, "COph100: A comprehensive fundus image registration dataset from infants constituting the “RIDIRP” database," *Scientific Data*, vol. 12, no. 1, pp. 99,2025 doi: 10.1038/s41597-025-04426-w.

[53] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, S. Douma, and A. A. Argyros, "FIRE: Fundus Image Registration dataset," *Modeling and Artificial Intelligence in Ophthalmology*, vol. 1, no. 4, pp. 16-28,2017 doi: 10.35119/maio.v1i4.42.

[54] D. Rivas-Villar, Á. Hervella, J. Rouco, and J. Novo, "Joint keypoint detection and description network for color fundus image registration," *Quantitative Imaging in Medicine and Surgery*, vol. 13, no. pp. 4540-4562,2023 doi: 10.21037/qims-23-4.

[55] 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, pp. 381-395,1981 doi: 10.1145/358669.358692.

[56] D. Rivas-Villar, A. R. Motschi, M. Pircher, C. K. Hitzenberger, M. Schranz, P. K. Roberts*, et al.*, "Automated inter-device 3D OCT image registration using deep learning and retinal layer segmentation," *Biomedical Optics Express*, vol. 14, no. 7, pp. 3726-3747,2023 doi: 10.1364/BOE.493047.

[57] X. Feng and G. Cai, "Retinal Mosaicking with Vascular Bifurcations Detected on Vessel Mask by a Convolutional Network," vol. 2020, no. pp. 7156408,2020 doi: 10.1155/2020/7156408.

[58] S. Mukherjee, T. De Silva, P. Grisso, H. Wiley, and D. L. K. Tiarnan, "Retinal layer segmentation in optical coherence tomography (OCT) using a 3D deep-convolutional regression network for patients with age-related macular degeneration," vol. 13, no. 6, pp. 3195-3210,2022 doi: 10.1364/boe.450193.

[59] K. Akyol and B. Şen, "Keypoint detectors and texture analysis based comprehensive comparison in different color spaces for automatic detection of the optic disc in retinal fundus images," *SN Applied Sciences*, vol. 3, no. 9, pp. 774,2021 doi: 10.1007/s42452-021-04754-7.

[60] A. Sharma, J. D. Oakley, J. C. Schiffman, D. L. Budenz, and D. R. Anderson, "Comparison of automated analysis of Cirrus HD OCT spectral-domain optical coherence tomography with stereo photographs of the optic disc," *Ophthalmology*, vol. 118, no. 7, pp. 1348-57,2011 doi: 10.1016/j.ophtha.2010.12.008.

[61] W. Wang, D. A. Miller, H. B. Price, X. Yang, W. J. Brown, and A. Wax, "High-Performance, Low-Cost Optical Coherence Tomography System Using a Jetson Orin Nano for Real-Time Control and Image Processing," *Transl Vis Sci Technol*, vol. 14, no. 3, pp. 24,2025 doi: 10.1167/tvst.14.3.24.

[62] J. Flusser and T. Suk, "A moment-based approach to registration of images with affine geometric distortion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 2, pp. 382-387,1994 doi: 10.1109/36.295052.

[63] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2D2: Reliable and Repeatable Detector and Descriptor," ed: NAVER LABS Europe.

[64] A. Ultsch and J. Lötsch, "Euclidean distance-optimized data transformation for cluster analysis in biomedical data (EDOtrans)," *BMC Bioinformatics*, vol. 23, no. 1, pp. 233,2022 doi: 10.1186/s12859-022-04769-w.

[65] J. M. Martínez-Otzeta, I. Rodríguez-Moreno, I. Mendialdua, and B. Sierra, "RANSAC for Robotic Applications: A Survey," *Sensors*, vol. 23, no. 1, pp. 327,2023 doi:

[66] C. A. N. Santos and N. D. A. Mascarenhas, "Patch similarity in ultrasound images with hypothesis testing and stochastic distances," *Computerized Medical Imaging and Graphics*, vol. 74, no. pp. 37-48,2019 doi: https://doi.org/10.1016/j.compmedimag.2019.03.001.

[67] C. A. N. Santos and N. D. A. Mascarenhas, "Geodesic Distances in Probabilistic Spaces for Patch-Based Ultrasound Image Processing," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 216-226,2019 doi: 10.1109/TIP.2018.2866705.

[68] C. A. N. Santos, D. L. N. Martins, and N. D. A. Mascarenhas, "Ultrasound Image Despeckling Using Stochastic Distance-Based BM3D," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2632-2643,2017 doi: 10.1109/TIP.2017.2685339.

[69] J. Schottenhamml, T. Würfl, S. B. Ploner, L. Husvogt, B. Hohberger, J. G. Fujimoto*, et al.*, "SSN2V: unsupervised OCT denoising using speckle split," *Scientific Reports*, vol. 13, no. 1, pp. 10382,2023 doi: 10.1038/s41598-023-37324-5.

[70] R. Panigrahy, "An Improved Algorithm Finding Nearest Neighbor Using Kd-trees," Berlin, Heidelberg, 2008, pp. 387-398.

[71] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An algorithm for finding best matches in logarithmic expected time," *ACM Transactions on Mathematical Software (TOMS)*, vol. 3, no. 3, pp. 209-226,1977 doi: