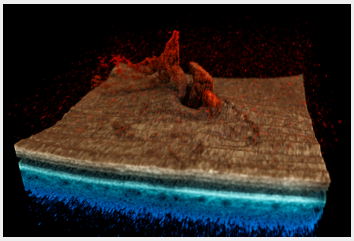
Computer Aided Medical Procedures
Prof. Dr. Nassir Navab



Dissertation

Intraoperative Visualization for OCT-Assisted Retinal Microsurgery

Jakob Matthias Weiss

Fakultät für Informatik
Technische Universität München

Technische Universität München
Fakultät für Informatik

Intraoperative Visualization for OCT-Assisted Retinal Microsurgery

Jakob Matthias Weiss

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende(r): Prof. Dr. Nils Thuerey

Prüfer der Dissertation: 1. Prof. Dr. Nassir Navab

2. Prof. Dr. Bernhard Preim

Die Dissertation wurde am 12.11.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 04.04.2022 angenommen.

Jakob Matthias Weiss
Intraoperative Visualization for OCT-Assisted Retinal Microsurgery
Dissertation, Version 1.02

Technische Universität München
Fakultät für Informatik

Boltzmannstraße 3
85748 and Garching bei München

Abstract

Intraoperative imaging is one of the cornerstones of modern interventional procedures. It is particularly important for ophthalmic microsurgery where delicate structures must be manipulated in the submillimeter range and small errors can have serious consequences. Retinal microsurgery is a particularly challenging field due to its restricted endoscopic access pathway, limited optical feedback and lack of haptic feedback. Surgeons commonly require many years of experience to perform advanced retinal procedures such as membrane peeling or subretinal injections.

In recent years, three technologies have emerged that, when combined, will transform microsurgical practice in ophthalmology: Binocular surgical microscopes are equipped with high-quality stereoscopic cameras to create digital microscopes that can be viewed on stereoscopic screens. Optical coherence tomography (OCT) has been integrated into the surgical microscope to provide high-quality cross-sectional images of retinal tissue and surgical instruments in real time. The latest OCT technology now also provides real-time 3D images of retinal tissue, greatly increasing the level of detail. Finally, novel robotic micromanipulators developed specifically for ophthalmic surgery promise more precise and controlled manipulation, overcoming the limitations of manual surgery and enabling procedures previously considered infeasible. This combination of novel technologies in conjunction with challenging clinical requirements creates a unique environment for visualization research. The quality of intraoperative visualization must match the precision of the robotic system to fully exploit the potential of robotic surgery.

This dissertation explores how the real-time information provided by OCT can be used to create intraoperative visualizations that support retinal surgery. It outlines a robotic surgical workflow that combines perioperative visual planning and live visual feedback during the manipulation. It also presents a concept for a future surgical environment based on virtual reality (VR) technology that takes advantage of the perceptual and interaction benefits of such immersive systems to further enhance the visualizations. Further contributions to this goal are presented: As a basis for visual feedback, an OCT-based instrument tracking algorithm provides precise tracking at low computational cost. OCT visualization modes are introduced that are specifically designed to exploit the layered structure of the retinal tissue to provide clear intraoperative visualizations. This is extended into an efficient pipeline for real-time volumetric OCT imaging data that aggregates volume data over time to improve image quality and enable enhanced visualization. Finally, a novel volume rendering method called Deep Direct Volume Rendering (Deep DVR) is presented. It leverages a tight integration of DVR with machine learning to learn semantic rendering purely from example images. This obviates the complex fine-tuning required for traditional DVR visualizations and has potential applications beyond ophthalmic imaging.

Zusammenfassung

Die intraoperative Bildgebung ist einer der Eckpfeiler der modernen interventionellen Verfahren. Sie ist besonders wichtig für die ophthalmologische Mikrochirurgie, wo empfindliche Strukturen im Submillimeterbereich manipuliert werden müssen und kleine Fehler schwerwiegende Folgen haben können. Die Netzhautchirurgie ist ein besonders anspruchsvolles Gebiet, da der endoskopische Zugangsweg beschränkt ist, das optische Feedback begrenzt ist und haptische Rückmeldungen fehlen. Chirurgen benötigen in der Regel viele Jahre Erfahrung, um fortgeschrittene Eingriffe an der Netzhaut wie Membran-Peeling oder subretinale Injektionen durchzuführen.

In den letzten Jahren sind drei Technologien etabliert, die in ihrer Kombination die mikrochirurgische Praxis in der Augenheilkunde verändern werden: Binokulare Operationsmikroskope werden mit hochwertigen stereoskopischen Kameras ausgestattet, um zusammen mit stereoskopischen Bildschirmen digitale Mikroskope zu schaffen. Die optische Kohärenztomographie (OCT) wurde in das Operationsmikroskop integriert, um hochwertige Querschnittsbilder von Netzhautgewebe und chirurgischen Instrumenten in Echtzeit zu liefern. Die neueste OCT-Technologie liefert nun auch 3D-Bilder des Netzhautgewebes in Echtzeit, was die Detailgenauigkeit deutlich erhöht. Schließlich versprechen neuartige, speziell für die Augenchirurgie entwickelte robotische Mikromanipulatoren eine präzisere und kontrollierte Manipulation, die die Grenzen der manuellen Chirurgie überwindet und Verfahren ermöglicht, die bisher als undurchführbar galten. Diese Kombination neuartiger Technologien in Verbindung mit anspruchsvollen klinischen Anforderungen bildet ein einzigartiges Umfeld für die Visualisierungsforschung. Die Qualität der intraoperativen Visualisierung muss der Präzision des Robotersystems entsprechen, um das Potenzial der Roboterchirurgie voll auszuschöpfen.

In dieser Dissertation wird untersucht, wie die von OCT bereitgestellten Echtzeitinformationen zur Erstellung intraoperativer Visualisierungen genutzt werden können, die die Netzhautchirurgie unterstützen. Es wird ein robotergestützter chirurgischer Arbeitsablauf skizziert, der perioperative visuelle Planung und visuelles Live-Feedback während der Manipulation kombiniert. Außerdem wird ein Konzept für eine künftige chirurgische Umgebung vorgestellt, die auf der Technologie der virtuellen Realität (VR) basiert und die Wahrnehmungs- und Interaktionsvorteile solcher immersiven Systeme nutzt, um die Visualisierung weiter zu verbessern. Weitere Beiträge zu diesem Ziel werden vorgestellt: Als Grundlage für das visuelle Feedback bietet ein OCT-basierter Tracking-Algorithmus für Instrumente eine präzise Verfolgung bei geringen Rechenkosten. Es werden OCT-Visualisierungsmodi eingeführt, die speziell darauf ausgelegt sind, die Schichtstruktur des Netzhautgewebes auszunutzen, um klare intraoperative Darstellungen zu liefern. Dies wird zu einer effizienten Pipeline für volumetrische OCT-Bildgebungsdaten in Echtzeit erweitert, die Volumendaten über die Zeit aggregiert, um die Bildqualität zu verbessern und eine verbesserte Visualisierung zu ermöglichen. Schließlich

wird eine neuartige Volumenrendering-Methode namens Deep Direct Volume Rendering (Deep DVR) vorgestellt. Es nutzt eine enge Integration von DVR mit maschinellem Lernen, um die semantisches Darstellung allein aus Beispielbildern zu lernen. Dadurch entfällt die komplexe Feinabstimmung, die für herkömmliche DVR-Visualisierungen erforderlich ist. Daraus ergeben sich auch potenzielle Anwendungsmöglichkeiten über die augenmedizinische Bildgebung hinaus.

Acknowledgments

These last couple of years have been an exciting journey for me, yet I would never have made it this far without the amazingly supportive people that shared this journey with me.

I would like to thank my advisor, Prof. Dr. Nassir Navab: Your guidance, encouragement and unique mentorship have continuously pushed me to pursue the research directions I was genuinely interested. Thank you for granting me this freedom, and for creating this great environment that allows people to grow in many more ways than just as a researcher.

The entire CAMP group is composed of many great individuals, yet some have shaped my path early on: Christian Schulte zu Berge, Ulrich Eck and Nikola Rieke were invaluable mentors of different aspects of my academic journey, and I want to thank you for the advice and support throughout the years. It has been truly inspirational to work with so many incredibly talented colleagues, and it was an even greater pleasure to find that so many are simply great people that are easy to become friends with. Thank you Maximilian Baust, Tobias Lasser, Federico Tombari, Shadi Albarqouni, Salvatore Virga, Iro Laina, Christian Rupprecht, Fabian Manhard, Hessam Roodaki, Rüdiger Göbl, Julia Rackerseder, Beatrice Demiray, Javier Esteban, Maria Tirindelli. Particular thanks to Alexander Winkler, Felix Bork, Matthias Grimm and Matthias Seibold, and especially Alejandro Martín-Gomez with whom I shared many long days and nights at the NARVIS lab. A special thank you goes to Ari Tran - thank you for being the support that carried me through much of this way, for all the times you provided encouragement, distraction, focus or chocolate at the right moments to keep me going.

Lastly, a big thank you to my brothers and especially my parents - I am deeply grateful for your love and continued support through so many years of university. Your kindness and understanding patience have always kept me grounded.

Thank you all.

Contents

Part I

Introduction and Background

1 Introduction and Motivation

Our modern world is predominantly built around perceiving information with our eyes: reading text or symbols like street signs, watching movies or visual arts, or simply gleaming the emotional state of another person from their facial expression: we are constantly processing and evaluating visual information throughout the day. Of the five primary senses, vision is widely considered the most important sense in our daily lives, yet we rarely stop to consider its importance unless our visual sense is impaired in some way. However with an aging global population, age-related visual impairments are becoming ever more prevalent. While many of these impairments like presbyopia or cataract can be treated easily effective treatments for retinal diseases are needed more than ever.

This dissertation considers *vision* from two very different angles: in a *perceptual* sense, it attempts to find effective ways to visually convey complex imaging data. This combines modern understanding of human visual perception and scientific visualization in order to display real-time information to a surgeon in a way that facilitates intra-operative decision making. However, within the context of retinal microsurgery, there is also an *clinical* aspect: in developing methods to directly support ophthalmic surgeries, the eye itself as the visual sensory organ becomes the object of manipulation. Both anatomical and pathological features of the eye, and in particular the retina, are critical to understanding which structures are contained in the imaging data and which details need to be made salient during a surgery. Equally important to the development of surgical visualization is an understanding of the technology involved in modern ophthalmic microsurgery. We will therefore begin by briefly reviewing retinal microsurgery technology from a historical perspective and then proceed to outline our motivations in that context.

1.1 Retinal Microsurgery Technology

The history of ophthalmic surgery literally spans thousands of years: historical evidence suggests that early versions of cataract surgery have been practiced as early as 2467-2457 BC in ancient Egypt [4]. Depictions and reports of cataract treatment using comparably primitive methods can be found in many prehistoric civilizations through the ages, yet what is notable is that most of these more than four millenia of opthalmic surgery have been exclusively concerned with treating the anterior section of the eye. In contrast, retinal surgeries have only really been performed for about 100 years [4], and reliable surgical techniques are even more recent: Even though the Japanese researchers T. Dodo and C. Haruta already published vitrectomy techniques in the 1950s [205], the birth of modern vitreous surgery (at least in the western world) is widely attributed to the influential work of Robert Machemer starting from 1971 [134]. His invention of the Vitreous Infusion Suction Cutter (VISC), a device that can simultaneously cut up and remove the gelatinous vitreous humour inside the

eye and directly replace it with liquid from an infusion line, and the subsequent refinement of microsurgical techniques in general, profoundly shaped the way retinal microsurgery is performed today. Several technological advances have influenced the current clinical practice of vitreoretinal surgery over the years, but another major turning point was the binocular indirect ophthalmomicroscope (BIOM), a surgical microscope with a special lens system that allowed for direct wide angle view of the fundus (the view onto the retina) without a contact lens on top of the eye held in place by an assistant. This technology is still the basis of modern ophthalmic surgical microscopes.

Another game-changing technology was adopted relatively recently: Optical Coherence Tomography (OCT), an imaging modality that is based on measuring backscattered light to create 2D images of tissue cross-sections, has been adopted quickly after its first use in retinal imaging in 1991 and has turned into the de-facto standard imaging method for diagnosing and managing retinal diseases. Its high resolution and non-ionizing, non-contact imaging mechanism make it an ideal candidate not only for diagnostic imaging but also for use during retinal surgery. Initial successful intraoperative use of handheld devices was followed by integration directly into the surgical microscope [51] in 2011, an approach that was translated to the OR exceptionally quickly with the first device receiving FDA approval only three years later [50]. Continuing research in OCT imaging has brought significant improvements in quality and imaging speed. Novel SS-OCT (Swept Source OCT) systems use frequency-tunable infrared lasers which are now able to perform volumetric imaging intraoperatively at high volume rates[111], a technology quickly dubbed 4D OCT (3D + time). This greatly increases the amount of detail that can is captured by the OCT imaging and allows for unprecedented detail in image guidance during retinal surgery.

In parallel, microscope manufacturers have begun producing fully digital microscopes by replacing the view through the microscope eyepiece by high quality stereo cameras. The imaging information is instead displayed on high resolution stereoscopic screens placed in the OR. In clinical practice, these *Heads-up 3D Surgery* systems provide image quality that is comparable to the conventional binocular systems with the added benefits of superior ergonomics and a view of equal quality for all observers in the OR [205]. This might seem like minor improvements, however it will be an important enabler technology for further digital improvements in future systems. First studies on multi-modal integration [57] between digital microscopes and intraoperative OCT have already shown positive results.

Lastly, groups at multiple research centers have been working on surgical robots especially designed for ophthalmic microsurgery [210]. These robotic systems offer superior stability and precise movement compared to manual manipulation of the instruments. The robotic system built by Preceyes B.V. is the first commercial system to receive CE certification in 2019. It provides a positioning accuracy of $< 20\,\mu\mathrm{m}$ and features like motion scaling and active tremor filtering [247]. Robotic manipulators in general will enable safer execution of existing treatments and pave the way towards novel surgical techniques.

Image source: [4] ©2009 Wolters Kluwer Health, Inc

(a) Egyptian Relief, ca. 1200 BC

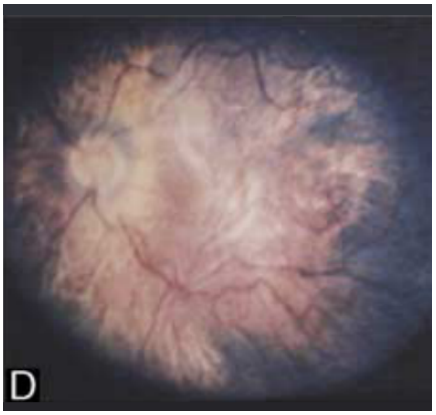


Image source: [205] ©2020 Wolters Kluwer Health, Inc

(b) BIOM, 1997

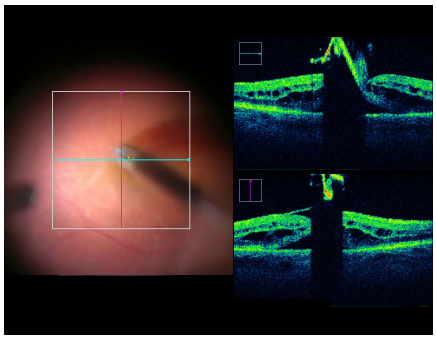


Image source: Courtesy Dr. Mathias Maier.

(c) Intraoperative OCT, 2014

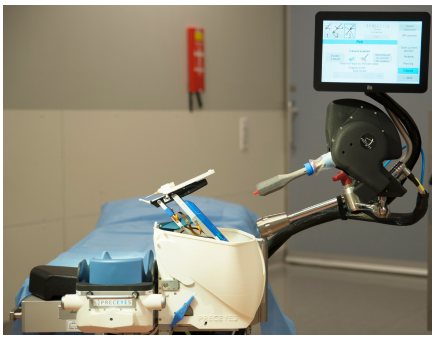


Image source: ©[247] 2019 Preceyes B.V.

(d) PRECEYES Surgical System, 2018

Fig. 1.1. Retinal Microsurgery Technology: From the primitive tools used in prehistoric times **(a)**, better intra-operative guidance through contactless retinal imaging **(b)** and intraoperative OCT imaging **(c)** has transformed ophthalmic surgery procedures. Novel robotic systems **(d)** promise unprecedented surgical accuracy for ophthalmic microsurgery.

1.2 Motivation

The drastic technological advancements of the last 50 years highlight the fast pace at which surgical technology proceeds, and the opportunities promised by the technology developed in the last decade are already met with much enthusiasm by the clinical community. The fast adoption rate of new technology is driven by the prevailing needs for better visual feedback to support the intricate surgical maneuvers.

The combination of advanced 3D imaging, digital microscopy and surgical robotics will have a lasting transformative effec, yet each of the technologies is currently limited when used in isolation: Intraoperative SS-OCT needs to be rendered on a digital screen, yet adding an external screen into the OR requires surgeons to divide their attention between the direct view through the microscope ocular and the screen with OCT imaging. Semi-transparent displays integrated into the oculars of some microscopes struggle with limited contrast and color reproduction. *Heads up 3D surgery* systems by themselves offer only small benefits that

make it hard to justify increased cost and more complex logistics of adding a large screen to already crowded surgical theaters. The high positioning accuracy of robotic manipulators can only be exploited when the operator can actually confirm the positioning with the same precision, which is hard with the limited fundus view offered by optical microscopes. Due to the limited availability of each of the technologies, research centers are only slowly starting to combine these technologies to determine these potential synergies.

An important factor in the advancement of precision surgery is the development of novel gene therapy vectors [62, 161, 196], which could not only halt degenerative retinal disease progression but potentially even restore lost photoreceptors through the injection of stem cells. The highly specific nature of these novel therapeutic agents results in very high costs per dose. For example Luxturna (Spark Therapeutics, Inc., Pennsylvania, USA), a recently approved gene therapy agent that can repair a specific genetic defect and potentially restore sight to the congenitally blind, is priced at 850.000 USD for a single treatment of both eyes [243]. Precise delivery of the agent to the affected region can maximize their therapeutic effect and could thereby reduce the required doses or at least improve the chances of a successful outcome.

Yet not only these highly specific, small patient populations will benefit from more precise and safe retinal interventions: The prevalence of age-related macular degeneration is currently estimated at approximately 67 million people in the EU, and it is expected to grow by 15% until 2050 due to population ageing[121, 176]. Worldwide, it is estimated that by 2050 288 million people will be affected by AMD [227]. Therapeutic agents currently in development to treat AMD will equally benefit from increased efficacy when administered subretinally [161, 228] and in light of this high prevalence, developing safe and effective methods to administer these treatments is an important goal.

1.3 Thesis Objective

These considerations highlight the great potential that the combination of digital microscopes, intraoperative OCT and robotic manipulation have to extend the operator's capabilities beyond human limits: robotic manipulation enables micrometer-precise manipulation, intraoperative 4D OCT provides high-resolution views into the tissue and digital microscopes pave the way for digital enhancement and composition of all information. Future advances in robotic technology will undoubtetly be necessary as well, but we believe that effective image guidance through intraoperative visualization of OCT data will be decisive for the success of both SS-OCT guided interventions and robotic treatment. Despite the unique challenges, this specific modality has so far only received little attention from the medical image computing and visualization communities. This thesis is therefore dedicated to find novel ways to present this imaging information intraoperatively. In short, the objective of this thesis is to find answers to the following question:

> How can intraoperative 4D OCT be effectively visualized in the operating theater to improve a surgeons' understanding of the surgical region of interest during conventional or robotic surgery?

1.4 Outline

This thesis is structured in four parts. **Part I** outlines the general motivation of this thesis and provides background information. **Chapter 2** explains the basics of retinal microsurgery and OCT technology. It also introduces the basic theory and practice of direct volume rendering, an elementary algorithm for 3D OCT visualization.

Part II outlines a system that combines OCT imaging with a teleoperated robotic platform and intraoperative visualization. This part presents proof of concept implementations that explore the potential of such an integrated system and identify its current technical limitations.

Chapter 3 explains the surgical environment of an integrated operating room and presents a modified workflow for visually assisted robotic subretinal injections.

Chapter 4 shows a combined visualization system for perioperative planning and intraoperative visualization. The setup is built on top a prototype heads-up display system and demonstrates how OCT and digital microscope can be combined into an improved system for robotic interventions.

Chapter 5 expands on the previous prototype by describing an intrasurgical guidance concept embedded in an augmented virtuality environment. A VR headset is used to show the intraoperative data to the surgeon while they control the robot via a controller.

Chapter 6 summarizes the advances that are necessary to fully realize the integrated environment outlined in the previous chapters. These insights are used to frame the methodology presented in the subsequent chapters.

Part III presents the methodological contributions made to advance the state of the art towards such a system.

Chapter 7 first addresses the problem of instrument tracking by presenting a tracking algorithm that is purely based on OCT images. The approach uses geometric modeling of the instrument cross-section in OCT, which leads us to an efficient algorithm that can track instruments across subsequent B-scans. We demonstrate how the tracking information can be used for a camera-based AR guidance application.

Chapter 8 presents visualization concepts adapted for retinal intraoperative OCT. Based on the layered structure of the retinal tissue in OCT, two visualizations are developed: Layer-aware DVR is a 3D visualization specifically optimized to show superficial structures of the retina to assist membrane peeling procedures. Layer-Adjusted MIP (LA-MIP) is a projective visualization that compensates for the natural retinal curvature to provide a clear visualization during subretinal injection.

Chapter 9 extends this concept to a full intraoperative processing pipeline capable of processing 4D OCT in real time. It applies temporal registration and real-time segmentation to perform temporal filtering and OCT view extension using data from

previous frames. A dedicated visualization mode facilitates instrument visibility and takes care that the reliability of the displayed data is made apparent to the observer.

Chapter 10 takes a more general view on the problem of volume data visualization and proposes a generalized approach for learning-based volume rendering called *Deep DVR* based on a formulation using arbitrary latent color spaces. Motivated by an attempt to unify the classic DVR pipeline and replace explicit feature design and definition of multidimensional transfer functions, this work attempts to answer the question of whether learning-based rendering can be trained to detect implicit semantic and desired visual properties only from images.

Finally, **Part IV, chapter 11** concludes the thesis and provides an outlook on future directions beyond the work presented. The appendix (**Part V**) contains a list of publications, abstracts of publications not directly discussed in this thesis as well as lists of figures, tables and the bibliography.

Substantial parts of this thesis have already been published and the respective publications are indicated at the beginning of each chapter or section. Throughout this thesis, I will use the first person plural to highlight the fact that much of the works were indeed a collaborative effort of an often multidisciplinary team.

2 Background

Contents

This chapter will establish the relevant clinical and technical background relevant for the contributions in this thesis. We will first introduce the basics of retinal microsurgery to establish the clinical context. We will proceed to explain the foundations of OCT imaging as the basic imaging modality at the heart of the concepts introduced. The final essential technology is the volumetric rendering algorithm and its related concepts, which is at the core of several of the the visualization concepts in the proposed methods.

2.1 Retinal Microsurgery

Ophthalmic surgery is divided into anterior segment surgery and posterior segment surgery, depending on which part of the eye is operated on: Anterior surgery is mainly concerned with the outward-facing parts of the eye including the cornea, lens, iris, and ciliary bodies. Conversely, the posterior segment consists of the back two thirds of the eye and posterior surgery is concerned with structures at the interior of the eye (c.f. Figure 2.1).

2.1.1 Anatomy of the Eye

The eyeball is a slightly aspherical globe of approximately $24\,\mathrm{mm}$ diameter. Its structure is depicted in Figure 2.1. The protective outermost layer is formed by the opaque white *sclera* and the transparent *cornea*. The *uvea* forms the second layer and consists of the *iris*, *ciliary*

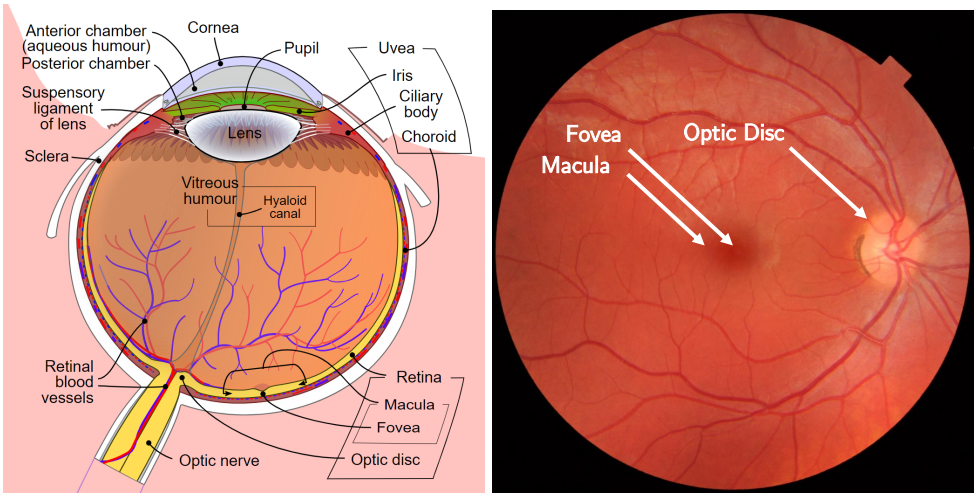


Image source: [99] CC BY SA

(a) Cross-sectional view of the eye

Image source: [79] CC 0, with modifications

(b) Fundus/Enface view of the retina

Fig. 2.1. Anatomy of the human eye. **(a)**: Schematic view of anterior and posterior structures in a coronal cross-section through the globe. **(b)**: An ophthalmoscope can be used to obtain a fundus view of the retina.

body and *choroid*. The innermost layer is the light-sensive *retina* and it covers about two thirds of the inside of the eyeball.

The interior of the eye is commonly divided into three parts: The *anterior chamber* is between cornea and iris. The iris controls the amount of light that enters the eye by relaxing or contracting, thus changing the size of the pupil. The *posterior chamber* is between iris and lens. The lens consists of elastic refractive tissue and can be deformed by the ciliary muscles to change the optical system to near accomodation or far accomodation. The third section is the *vitreous chamber* and it encompasses the vitreous humour and retinal tissue. The vitreous humour is a transparent gel-like fluid and its main function is to keep the spherical shape of the eye.

Retinal blood vessels spread out from the optic nerve to form a fine network of capillaries that provide nutrients to the retinal cells. In the *Enface* or *fundus* view shown in Figure 2.1b, important areas can be distinguished: The *optic disc*, sometimes called the "blind spot" due to its lack of photoreceptors, is the area where the optic nerve enters the eye. The *macula* is the central area of the retina. It measures around 5mm in diameter and processes all of the central visual field. At its center is the *fovea*, an indentation in the retina that contains the highest density of photoreceptors and thus forms the spot of highest visual acuity.

The retina is composed of several membranes and cell layers that form an intricate network that detects, processes and transmits the light (see Figure 2.2). In the context of this dissertation, the most important structures are: the *Inner Limiting Membrane* (ILM) is a thin ($5 - 10\,\mu\mathrm{m}$), collagen-rich membrane that separates vitreous humour from the neurosensory retina. The layers inbetween the ILM and the *External Limiting Membrane* (ELM) form the *sensory retina* and contain the photosensitive nerve receptors that detect incoming light intensity and color. The *retinal pigment epithelium* (RPE) is a pigmented cell layer that serves metabolic, nutritional

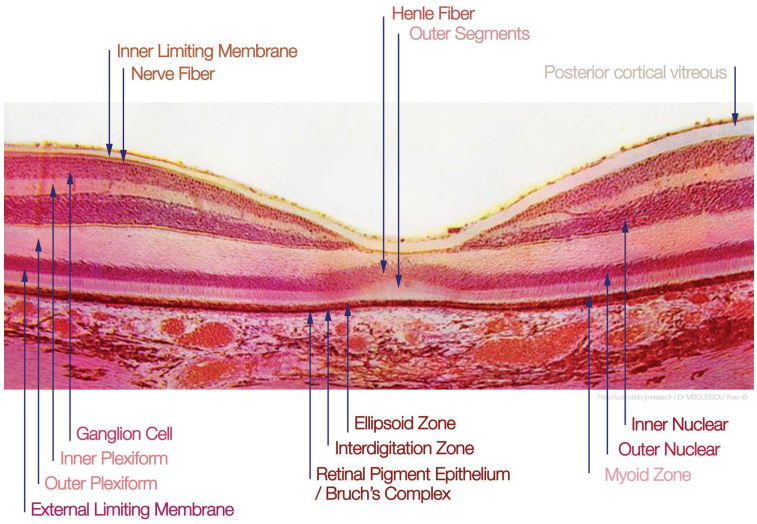


Image source: [141]

Fig. 2.2. Retinal Layers in Histology

and immunological functions [196] by communicating nutrients and other agents between the photoreceptors and the *choroid* located below.

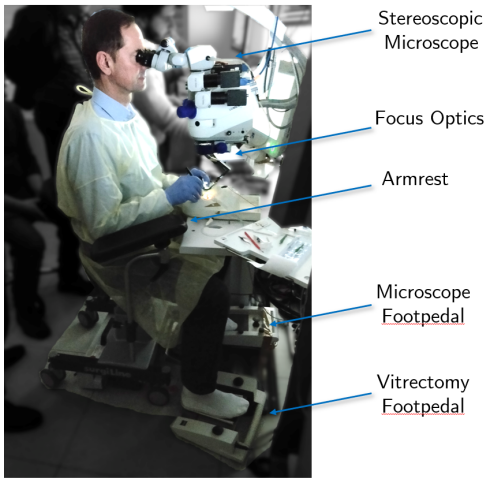
2.1.2 Vitreoretinal Surgery Setup

Ophthalmic surgery theaters are equipped with highly specialized instruments that allow to perform these surgeries in a minimally invasive manner. This specialized equipment and setup merits a closer look as it highlights the unique environment of vitreoretinal surgery compared to other kinds of surgery.

Generally the patient is supine on the operating table and the surgeon is sitting superior relative to the patient, essentially looking upside down onto the patient. Although in some cases, special positioning such as an upright posture might be indicated [190, 232], a supine patient is the conventional setup for surgery.

Surgical Access

The interior of the eye globe is usually accessed through the sclera. While surgical access via *sclerotomy* ports (incisions) was common practice up until the early 2000s, the introduction of micro-incision suture-less vitrectomy surgery (MIVS) in 2002 [151] largely superseded sclerotomy in routine practice. With MIVS, the vitreoretinal space is accessed via trocars through the sclera placed approximately 4mm from the border of the cornea [197], as seen in Figure 2.4b. This was shown to lead to "more efficient surgery, faster recovery time and better visual outcomes" [151]. For many surgeries MIVS has become standard practice, however 20G sclerotomy still has benefits in specific cases [17]. The trocars avoid excessive stress on the anatomy when moving or changing the instrument and additionally, through an integrated membrane, avoid leakage of vitreoretinal fluids. Instrument diameters are standardized



(a) General sitting configuration.

(b) Zeiss Lumera 700 with RESIGHT 700, Callisto touch-screen control and foot pedal

Fig. 2.3. Ophthalmic surgery setup. **(a)**: In a normal sitting position during (mock) surgery, the surgeon views the surgical site through the microscope and operates the endoilluminator light source with one hand and an active instrument with the other. Armrests mounted on the chair provide essential stability during the precise manipulations. In vitreoretinal surgery, the surgeon often uses two additional foot pedals: one controls the aspiration and cut rate during vitrectomy. The second controls the surgical microscope **(b)** and allows adjustment of microscope positioning and OCT parameters.

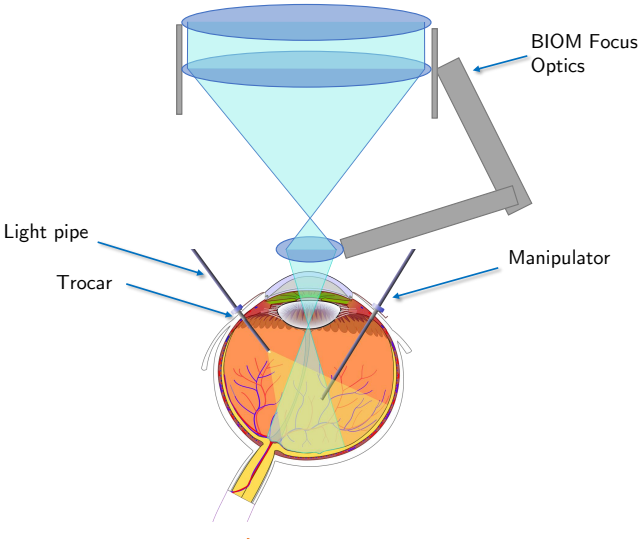
through the American Wire Gauge (AWG, see Table 2.1), and 23G or 25G instruments have been established standard sizes with 27G gathering recent interest [164].

In the standard setup, three trocars are placed (see Figure 2.4b), where each trocar has a specific function: the first one is placed inferonasally and is connected to the infusion line to stabilize intraocular pressure when vitreous or fluids are removed from the eye. The second and third are placed superonasally and superotemporally and are typically used for the endoilluminator (at the surgeon's non-dominant hand) and the active surgical instrument (i.e. forceps, vitrectomy cutter, injection needle).

Visual Access

Direct observation of retinal details is generally not possible without additional optics. A specialized ophthalmic surgical microscope is used to obtain visual feedback via the conventional optical pathway through the patients' dilated pupil. Much of the data acquisition and experiments in this dissertation has been performed based on the OPMI Lumera 700 (Carl Zeiss Meditec AG, Jena) ophthalmic microscope[23] platform. We will proceed describing this system as representative system, as most competing devices generally follow similar design principles.

As can be seen in Figure 2.3b, the binocular microscope is mounted on a movable mechanical arm so it can be brought directly above the patient eye at the start of the surgery. Separate BIOM optics, the RESIGHT 700, can be brought into the microscope's optical axis to invert the optical pathways of the patient's eye and enable contactless visualization of the patient's



(a) Visual and physical retinal access

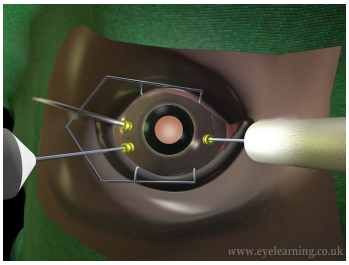


Image source: [197]

(b) Trocar Placement

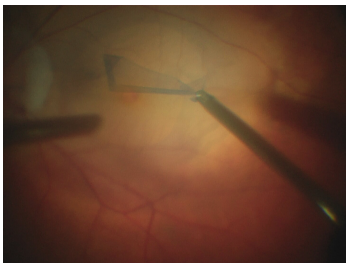


Image source: Courtesy Dr. Mathias Maier

(c) Intraoperative fundus view showing light source (left) and forceps (right).

Fig. 2.4. The visual pathway for retinal microsurgery uses special microscope optics to provide a view directly through the patient's pupil. To provide sufficient light inside the eye, an endoilluminator is introduced via a scleral trocar. Another trocar is used to introduce the active surgical instrument (*manipulator*), e.g. forceps, vitrectomy cutter or injection needle.

retina. During surgery, the view can be changed by selecting one of two different lenses: one provides a wide field of view and is used often for operations in the vitreous and the peripheral regions while the second lens provides a high magnification of the macula for precise manipulation in this critical space. The microscope has motorized focus and zoom capabilities to compensate for the patient-specific refractive properties and optimally adjust the microscope field of view to the surgical region of interest. Zoom and focus as well as lateral motion of the microscope can be digitally controlled either via buttons on the microscope handle or an external touchscreen monitor marketed as the CALLISTO platform. This external monitor also functions as a secondary display for the surgical scene, captured through a camera within the microscope, to allow spectators and surgical staff to follow the surgery. Additionally the surgeon can use a wireless foot pedal with six buttons and a two-axis joystick to acjust the surgical microscope. This foot pedal is essential during surgery to allow the operator to adapt the microscope zoom and focus as well as other integrated features as necessary without requiring their hands. Maintaining optimal visual conditions is challenging as the optical properties of the surgical site often change with a number of factors: changes in intraocular pressure or corneal shape caused by forces on the trocars lead to changed refractive properties of the anterior complex, which in turn leads to a defocused view. Rotation of the eye changes the visual axis and can induce vignetting, and replacement of the vitreous with other fluids or gases as part of the procedure dramatically changes the overall optical properties, leading to reflections and intermittent loss of visual feedback[197].

AWG	μm	Comment
20G	812	Used in sclerotomy surgery [151]
23G	573	MIVS [151]
25G	455	MIVS [151]
27G	361	Potential future MIVS systems [164]
30G	255	Intravitreal injections [196]
39-42G	63-90	Cannula of subretinal injection needles
	100-300	Thickness of human retina [119]
	40-350	Diameter of retinal vessels [119]
	17-180	Diameter of human hair [245]
	60	Epiretinal membrane thickness [225]
	5-10	ILM thickness [223]

Tab. 2.1. American Wire Gauge (AWG) values relevant for retinal surgery and corresponding metric values, together with other anatomical structures for comparison.

As an alternative to viewing the surgical scene through a microscope ocular, so-called 3D heads-up display systems have been introduced in recent years. These systems capture the microscopic view with two cameras to reproduce the image on a 3D screen inside the surgical theater to create a fully digital microscope. Commercial systems like the TrueVision®3D Visualization System (TrueVision Systems Inc., Santa Barbara, CA, USA) or the NGENUITY®3D Visualization System (Alcon, TX, USA) are available but have not yet seen widespread adoption. Clinical case studies report that these systems are on par with conventional microscopes as "evidence suggests that [they] provide similar surgical times, visual outcomes and complication rates" [148]. Furthermore, these reports highlight the high potential for teaching during live surgery as all observers can appreciate the surgeon's view with similar fidelity. Yet one of the biggest selling points is the superior ergonomics offered by these systems: neck and back strain is commonly experienced by ophthalmologists[143] and likely caused by the surgeon having to maintain a steady head position in front of the microscope ocular.

Illumination

Due to the light-absorbing nature of retinal tissue, active illumination is necessary for a clear view during surgery. The surgical microscope already has an integrated lightsource that is coupled into the optical pathway, which minimizes shadows from structures around the eye. However, the additional lenses of the microscope as well as the patient's anterior anatomy are prone to reflections, rendering the integrated light ineffective for retinal surgery. Instead, an endoilluminator probe is inserted into one of the trocars to provide light directly inside the ocular cavity. Endoilluminators consist of a fiber-optic light pipe that guides light from a dedicated external light source (typically *Xe* or halogen lamps) and disperses the light within a cone from the tip of the illuminator. Surgical illumination has however been reported as a potential cause for phototoxicity of retinal tissue[29, 66, 209]. Even though modern surgical light sources are equipped with special filters to mitigate the problem, surgeons are still required to consider phototoxicity by keeping the light source at a maximal distance from the retina and minimize exposure time. Additionally, the requirement of an endoilluminator means

that one of the surgeon's hand is occupied with handling the light source via one of the trocars. While surgeons reportedly use the dynamic control of the light source for improved depth perception by observing the instrument shadow on the retinal structures[210], a handheld illuminator prevents bimanual manipulation of tissue. *Chandelier* endoilluminators can be fixated inside the trocar which enables the surgeon to use a second instrument [17], however at the cost of a fixed, potentially suboptimal lighting and an additional scleral trocar.

2.1.3 Retinal Interventions

While many of the methods of this dissertation are applicable to a wider range of vitreoretinal procedures, they have been designed with two specific procedures in mind: membrane peeling and subretinal injection. Both of these methods have been identified in personal communications with surgeons as procedures where advanced imaging and visualization would be highly appreciated. In addition, they are good candidates for robotic surgery [210].

Epiretinal Membrane (ERM) and ILM Peeling

Epiretinal membrane, also known as *macular pucker*, is a pathological condition that consists of newly formed collagen and glial cells on the retinal surface. This tissue can tighten with age and, together with age-related shrinkage of the vitreous, can cause tension on the underlying tissue, often resulting in puckering or upward bulging. Blurred and distorted vision are common symptoms [246]. The prevalence of ERM rises with age from 1.9% at age 49-59 to as high as 11.6% at age 70-79 [147].

A pathology that is closely related to contraction of the vitreous is a *macular foramen*, or *macular hole*. This is usually caused by tension on the retinal tissue from the retreating vitreous gel [30] and it can cause blurred and distorted vision as well as blind spots. The prevalence has been reported as 3.3 per 1000 of the general population [61], again mainly affecting the older age groups.

In both of these cases, surgical treatment might be indicated in more severe cases. The first step of the surgery is perform a vitrectomy to remove the contracted vitreous and replace it with balanced saline solution (BSS), air, gas, or silicone oil to prevent further tension. Secondly, the ILM or ERM is peeled and removed, typically using a front-grasping forceps [30]. Removal of the membranes removes lateral surface traction and overall increases retinal elasticity [30], therefore promoting successful remodeling of the retina after surgery.

Removal of the membrane adherent to the surface, however, is a delicate procedure (see Figure 2.5). Finding a good starting point for pulling away the layer with a forceps relies on identifying an area where the membrane is more loosely attached to the retina, which is challenging to judge only from the fundus view through the microscope. Preoperative imaging can aid in identifying good candidate points, however mentally mapping the preoperative and intraoperative views remains a challenge. Once an initial part of the membrane is removed, the edge of the membrane simplifies grasping of the remainder. However, the pulling forces have to be moderated carefully to avoid complications such as retinal detachment or tearing, which

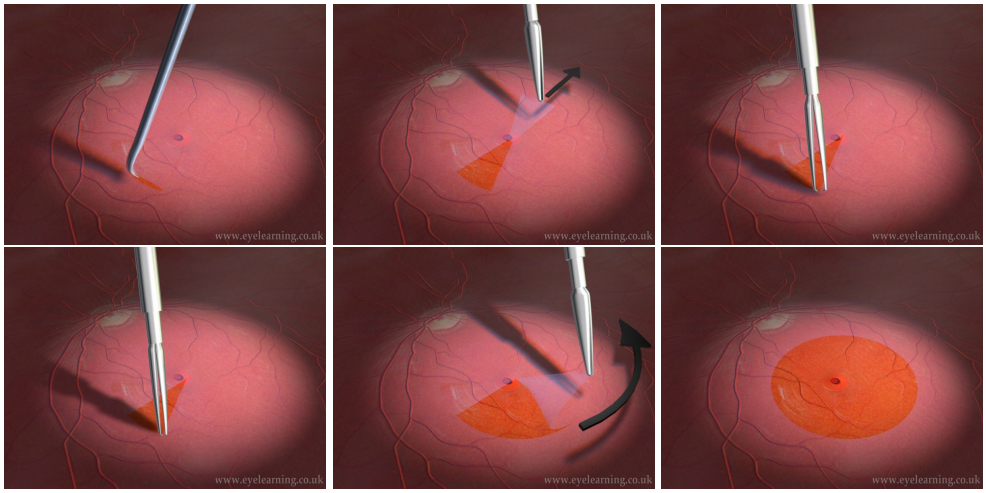


Image source: [197]

Fig. 2.5. Peeling procedure as suggested by Sullivan [197] to remove an ILM membrane that causes strain on the retinal tissue. After staining of the tissue with *brilliant blue* to increase visibility of the membrane, a tear is created with a small retinal pick or the forceps itself. Subsequently, a small strip is peeled off to create an edge that is easier to grasp. The flap is then torn away around the lesion to create a circular region of pliable retina around the foramen that facilitates closure.

are estimated to occur in 3-17% (detachment) and 1-5% (tears) of macular hole surgeries [65].

Subretinal Injection

Retinal degeneration, either through age-related factors (age-related macular degeneration, AMD) or through hereditary disorders (inherited retinal degeneration, IRD), often leads to partial or complete vision loss due to irreversible death of photoreceptors [136, 196]. AMD affects more than 10 million in the US alone[196] and it remains a major cause for blindness in the developed world. Recently, stem cell therapy has been gaining attention as potential treatments for these currently considered incurable diseases by being able to prevent or even reverse the loss of photoreceptors [136, 187]. While treatment varies in the type of stem cells used, and whether a regenerative or a trophic role is intended [244, 136], the general idea is to transplant stem cells into the retinal tissue as close as possible to the site of degeneration in order to heal damaged RPE cells or grow new ones. Gene replacement therapy, for example to prevent harmful neovascularization in the retina, faces the same challenges in delivery[196]. For both therapies, directly administering the agent subretinally, i.e. between the photoreceptor and RPE layers, is considered optimal from a therapeutic perspective when targeting the RPE [161, 196].

In general, delivery of therapeutic agents to the back of the eye is complicated by the anatomical barriers formed by the anterior complex. While topical delivery through eye drops is standard of care for many diseases of the anterior eye, it would not result in effective doses for the treatment of retinal diseases. The alternatives for treating vitreoretinal diseases are then either systemic delivery (which is not applied often due to potential systemic adverse effects [15]) and more targeted delivery through invasive procedures. Intravitreal injection is already often employed, for example for delivering anti-VEGF (vascular endothelial growth factor) drugs to the vitreoretinal space. This is however not as effective for treatments using

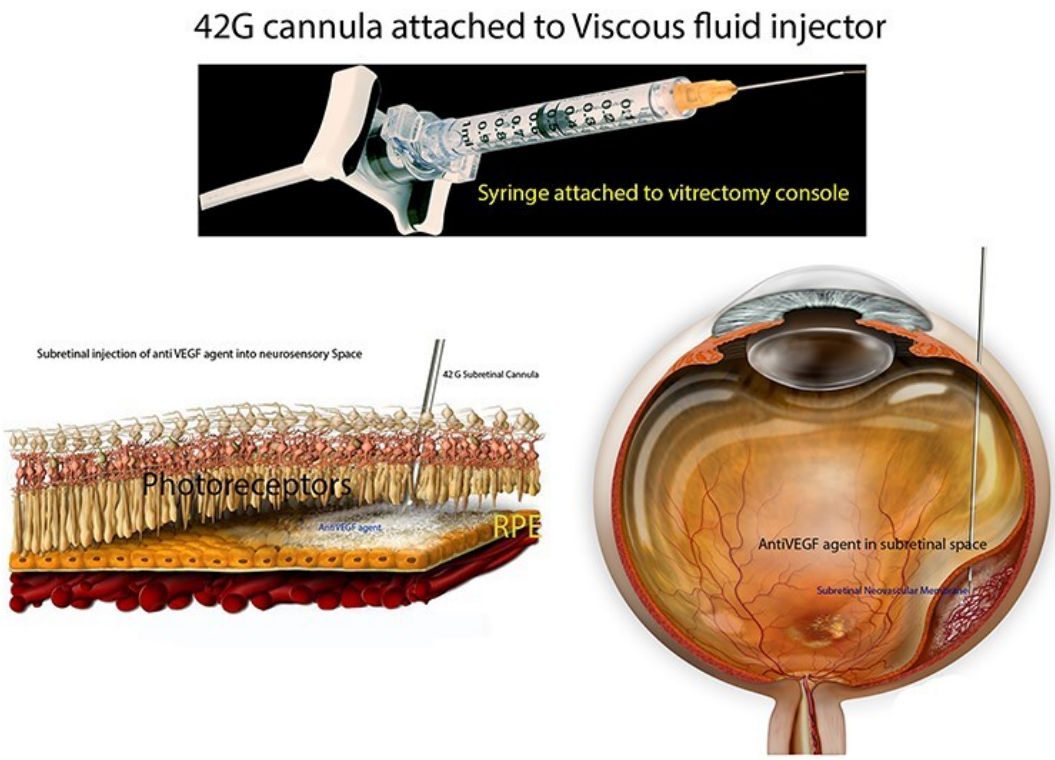


Image source: [28] ©2021 Oxford University Press

Fig. 2.6. Material and delivery pathway for the subretinal injection of aflibercept (Eylea; Bayer, Basel, Switzerland) [28]. A 25G injection needle with a 42G flexible injection tip was used to deliver about 0.05cc of the anti-VEGF agent into the subretinal space after subretinal hemorrhage. The retina was engaged at a 45-60deg angle to facilitate insertion and positioning was confirmed with iOCT. The authors reported "complete resolution of the submacular hemorrhage and good recovery of visual acuity".

gene and stem cell therapy [136, 187], where the ILM still presents a significant barrier for current treatment vectors [196].

Subretinal delivery is typically performed using the general surgical setup described above. Following a vitrectomy to avoid retinal detachment [196], an injection needle with a very thin (39-42G) cannula is used to penetrate the upper retinal layers and inject the agent directly in the subretinal space. An example of this procedure is depicted in Figure 2.6 for the delivery of an anti-VEGF drug into the subretinal space. Given the small size of the targeted area between the layers, injection demands very high surgical dexterity to avoid wrong placement of the bleb and prevent penetration of the RPE layer.

2.1.4 Assistive Technologies in Retinal Microsurgery

Ophthalmic surgeons need to deal with a number of major challenges when performing retinal surgery. Manipulation of the vitreoretinal space is restricted by the scleral ports. Since most surgeons use their dominant hand to manipulate the retina while the non-dominant hand controls the endoilluminator, access to a point on the retina is mostly limited to a single direction determined by the placement of the instrument port. Considering the microscopic scale of the structures, human limitations for precise manipulation become an important factor

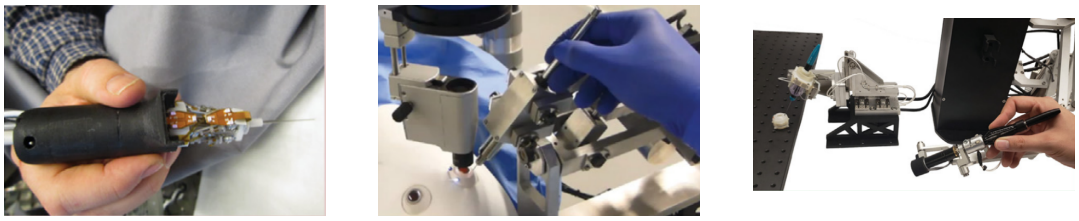
and it has been shown that natural hand tremor in these scenarios can be a limiting factor [82, 223]. Furthermore, for the majority of contact events in microsurgery, the generated forces cannot be perceived by the surgeon [76]. This lack of tactile feedback implies that surgeons have to rely heavily on visual information, yet the indirect view through the microscope is itself restricted. It is restricted by the pupil which implies a fixed viewing direction, complex, dynamically changing optical pathways as well as changing illumination. The axial fundus view also makes it difficult to accurately judge instrument positioning relative to the retina, especially when stereoscopic depth cues are lost at high magnification levels [197]. For subretinal injections, understanding the penetration depth from the fundus view directly becomes nearly impossible, requiring the surgeon to rely on their experience in interpreting subtle cues such as retinal deformation to place the cannula at the right depth. The difficult surgical conditions are compounded by the potentially disastrous consequences of wrong judgements such as retinal tears, detachment or hemorrhage which can ultimately lead to partial or complete vision loss.

This clinical demand for assistive systems has inspired the research and development of novel technology to aid these interventions. These have mainly focused on two different approaches: (1) improved intraoperative guidance and visualization and (2) improved instruments and manipulators to stabilize manipulation. The first approach aims to support surgical decision making by providing additional information and therefore extenuate the mentioned limitations of the surgical microscope. This direction has mainly focused on the use of intraoperative OCT, an imaging modality that can be integrated into the surgical microscope and provide cross-sectional images of the retina in real time. OCT will be discussed in detail in Section 2.2. The second direction of research in assistive systems is focused on mitigating the human limitations of precise manipulation through improved instruments. Mechanical and robotic solutions have been proposed to stabilize manual motion and perform specific surgical maneuvers. Due to their superior stability and precision, these systems can even enable procedures that were previously unfeasible. The next chapter will provide a summary of the current approaches.

2.1.5 Robotic microsurgery systems

The clinical need for robotic and assistive systems has been met with continuous contributions by the robotics community. Efforts to provide robotic stabilization have been developed as early as the teleoperated micromanipulator decribed by Ben Gayed et al. in 1987 [14, 75] and have matured towards in-human trials in 2018 [49] and first commercializations of such systems [247]. The key motivation of robotic systems for microsurery is the necessity to manipulate complex structures at scales that are at or below human possibilities in terms of dextrous manipulation, thus manual manipulation of the instruments becomes a limiting factor to perform these surgeries.

Robotic systems for ophthalmic microsurgery generally fall into four categories [228]: *hand-held* robotic solutions directly integrated with the instrument, *cooperative control* designs that hold the instrument together with the surgeon, *teleoperated* systems that are controlled via an input device that is not mechanically connected to the robot, and a fourth emerging category of *microrobots* that are dropped into the surgical area. Examples for the first three can be observed in Figure 2.7.



(a) Handheld [135] (b) Cooperative Control [70] (c) Teleoperated [152]

Image source: [210] ©2020 Elsevier Inc., with permission.

Fig. 2.7. Examples for the three major classes of microsurgical robots: the handheld *Micron* robot from Carnegie Mellon University **(a)**, the cooperative control system developed at KU Leuven **(b)** and the teleoperated micromanipulator developed at the University of Utah **(c)**.

Handheld systems integrate mechatronic components into a handheld form factor to introduce motion control between the surgeon's hand and the instrument tip [210]. In this design, the mechanical components can act as a buffer between the hand movement and the intended instrument movement. This way, instrument motion can be stabilized and unintended motion such as hand tremor can be filtered out. With such systems, the surgeon can perform the surgery with a similar posture and workflow as with non-assisted instruments while benefiting from the robotic stabilization.

A prominent example of this concept is the Micron system developed by a group at Carnegy Mellon University. Their design incorporates a 3DoF [135] and later 6DoF [229] piezoelectric actuator optimized for vitreoretinal surgery into an ergonomic handheld device. In integrating virtual fixtures generated in real time from the surgical scenes, they showed that motion scaling and tremor reduction can be an effective way to improve task performance in vein tracing and peeling tasks [12]. They showed the effectiveness of tremor reduction by 90% in a pointing task and a maximum error of $25\mu m$ in tracing tasks [229].

The SMART instruments developed at Johns Hopkins University present a different approach to a handheld design with only a single degree of freedom along the instrument axis. They integrate OCT into the instrument to acquire depth information along the instrument axis in real time. This feedback is then used to control the instruments' position along the single DoF in order stabilize the position of the instrument tip with respect to the retina. The group has demonstrated the feasibility for different surgical instruments such as a microforceps [193] and a system for subretinal injection [103].

Cooperative-control systems are characterized by the fact that the operator directly holds the instrument while the robot is also directly attached. Thus the surgeon still holds direct control and retains the familiar posture and movement while the robot can perform stabilizing and motion-locking features. Cooperative control is generally based on the robot's ability to perform force sensing on its actuators. In a form of admittance control, the controller software reacts to the implicit commands from the surgeon conveyed by pushing the instrument in a certain direction. This approach enables stabilization and motion filtering in the simplest case, but can be extended to restrict motion, for example to enforce (up to a point) distance or maximum force to the retina, or to allow only a single axis of motion during the injection to avoid retinal damage through lateral motion of the cannula.

The Steady Hand-Eye Robot (SHER) developed by the Johns Hopkins University [86, 87] was developed with such an admittance control system. In their design, they also incorporate a mechanical and a virtual remote center of motion (RCM) that allows rotational movements of the instrument only around a single point along the axis. This RCM point corresponds to the the position of the trocar and therefore lateral strain on the scleral port can be avoided when operating on the retina. This design approach is also followed by the robotic system presented by the Unversity College London [133]. In their design they focus on an additional degree of freedom provided by a rotational actuator at the base to quickly remove the robot from the working area when not in use. KU Leuven uses also presented a cooperative control platform aimed at retinal vein cannulation [69].

Teleoperated systems are divided into two subsystems responsible for the direct control of the robot and the input device operated by the surgeon. Sometimes called master-slave[1] systems, the two subsystems are not mechanically connected and therefore the robot is fully isolated from external forces of direct manipulation by a surgeon. The master device can then be freely chosen for an optimal user interface that suits the intervention as the control signals are only transmitted to the mechatronic component electronically. As input devices, haptic devices like the *Touch* (3D Systems GmbH, Moerfelden-Walldorf, Germany) are often chosen as they provide the necessary degrees of freedom and the pen-shaped handle has similar ergonomic properties as the standard ophthalmic instruments which facilitates mental mapping. When the slave device includes force measurement sensors, haptic feedback devices also enable force feedback scaling as a way to magnify the subtle forces to a perceptible level. Alternatives such as joysticks or 3D mice have been considered [5] and might be beneficial when the motion has to be mapped explicitly to specific princple directions. For example the forward motion during the critical phase of an injection might be better controlled with a single button or linear input device than through a full 6DoF input device.

The history of teleoperated systems for ophthalmology goes back to the stereotaxical microtelemanipulator for ocular surgery (SMOS) system of the Automatical Center of Lille [14, 75]. Many groups have developed teleoperated systems of diverse designs since then [63, 72, 83, 84, 94, 95, 208, 214, 226, 233], varying the mechanical design considerations and degrees of freedom, implementation of the RCM and electromechanical components [210].

A system of note is the Preceyes system, originally developed collaboratively at the TU Eindhoven and University of Amsterdam [142]. Their system, consisting of dedicated master and slave robots, is particularly well matured and the design also considers important practical challenges such as the mount to the operating table, compactness and the ease of installation and deployment. The system has since been commercialized by Preceyes B.V. [247] and was the first system to undergo in-human evaluation trials [49]. The Preceyes system has received a C.E. mark and is commercially available for use in vitreoretinal surgery in Europe [247]. Their instrument manipulator has four degrees of freedom with a controllable RCM point, can hold a range of different instruments and provides a precision below $20\mu m$. The motion controller has four degrees of freedom and is paired with an foot control pedal and a touch screen interface for additional inputs.

[1]The author acknowledges the problematic socio-historic implications of this terminology, however it is used here for clarity of exposure and consistency with the literature

The second system to highlight here is the iRAM!S system developed at the Technical University of Munich [153, 237], as this system was used as a reference in the concepts of the following chapters. The system uses piezoelectric motors as actuators in a parallel-serial arrangement which allows for a relatively compact design and offers 5DoF with an RCM implemented in software. In its current iteration, input devices ranging from a conventional joystick to a 3D mouse and a haptic feedback device [9] have been realized. The software-defined RCM is particularly useful when fitting the needle through the trocar when no RCM is needed, so it can be enabled once the needle tip reaches the posterior space. In their recent iteration [237], the system has been updated with the long-term goal to integrate OCT tightly into the robot control loop for improved instrument tracking and ultimately use information from OCT for closed-loop autonomous image-guided robotic interventions.

Microrobots. The use of untethered devices that can maneuver wirelessly from the sclera to the retina is a relatively novel concept that emerged only recently as a potential alternative to electromechanical manipulators. These microscopic actuators are not independent mechatronic devices but rather specially engineered, relatively simple devices that are controlled fully by an external magnetic system [116]. These devices can have either have tubular [31], ellipsoid [60] or a screw shape [137] and can be useful for delivery of small doses of potent drugs or for delicate operations such as retinal vein cannulation to dissolve occlusions. Targeted drug delivery is performed by coating the microrobot surface and subsequent diffusion into the tissue [60], a process that cannot be used when larger volumes of liquid have to be deployed such as in the case of gene or stem cell therapy. Therefore, these devices are currently not as versatile as the electromechanical alternatives described above.

2.2 Optical Coherence Tomography

Optical Coherence Tomography is an interferometry-based optical imaging modality that provides depth imaging into tissue at high spatial resolution. It uses non-ionising radiation, making it a safe imaging modality useful in a number of applications including cardiology and vascular imaging, endoscopic imaging, neurosciences as well as anterior and retinal ophthalmic imaging[46]. OCT has seen widespread adoption in ophthalmic applications, owing its success to the fact that internal tissue can be imaged non-invasively using the natural optical pathways of the eye and no competing imaging modalities exist that can provide similar resolution. Compared to ultrasound as the closest available alternative used in ophthalmic imaging, OCT provides 10-100 times higher axial resolutions of $1 - 15\mu m$ [45]. This imaging resolution enables imaging of the intricate microstructure of the retina as *in-vivo* optical biopsy.

OCT is nowadays tightly integrated into clinical routine where it is used to diagnose and manage a variety of retinal diseases like AMD, macular pucker, retinal detachments or subretinal fluid buildups. Since its first introduction in 1991 [93], OCT technology has developed rapidly. New light sources, scanning modes and detectors have progressed the technology from the first in-vivo imaging of the retina in 1993 [64, 200] to the first commercial diagnostic system in 1996 by Carl Zeiss Meditec, and is now standard of care in ophthalmology and optometry. Beyond diagnostic imaging, OCT has the advantage of relatively simple integration into other medical instruments as optical fibers can be used to deliver the sampling beam

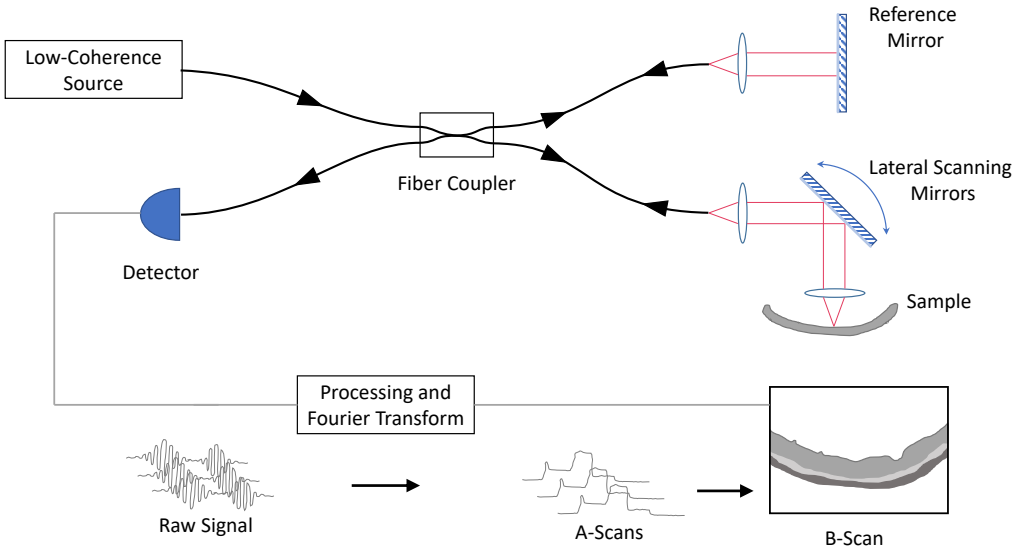


Fig. 2.8. Schematic of spectral/Fourier domain OCT imaging.

to the tissue. For retinal microsurgery, it can be integrated into the surgical microscope where it can partially share the same optical pathway to provide real-time cross-sectional imaging of the retina during a surgery. Advances in OCT imaging technology now enable the real-time intraoperative imaging with cross-sectional and even volumetric images. OCT is therefore of prime interest for image-guided opthalmic surgery [55, 210] for providing real-time high-resolution spatial feedback from the site of operation.

In the following, we will provide background on the general imaging process of OCT (Sec. 2.2.1). Section 2.2.2 will then introduce intraoperative OCT and discuss the factors that influence image quality.

2.2.1 OCT Imaging Technology

OCT is a tissue imaging modality that measures the light reflected or backscattered from different tissue depths. The imaging mechanisms of OCT are similar to ultrasound, in which sound waves are emitted into the tissue and image formation is performed from the echo delay and amplitude of detected backscattered sound. Instead of sound waves, OCT uses the infrared light of a light source with a typical center wavelength in the near infrared range (840-1310nm)[26], which enables much higher imaging resolutions than even high-frequency ultrasound. This frequency range is generally eye-safe when the power of the imaging beam is controlled. The much higher propagation velocity of light makes direct measurement of echo delay impossible, which is the reason why an interferometric approaches is used. The three principle variants for OCT imaging are Time Domain (TD), Spectral Domain (SD) and Swept-Source (SS) OCT, all of which rely on the principle of low coherence interferometry. The general setup follows a Michelson interferometer design and is depicted in Figure 2.8. The light from the light source is directed into a beam splitter, which can be an optical fiber coupler or a semi-transparent mirror. One part of the light goes through the so-called reference arm to be reflected at a reference mirror. The second part of the light is directed through the

sample arm towards the tissue to be imaged. The light reflected from the reference mirror is recombined with the light backscattered from the sample and fed into a detector. Light reflected at different depths of the sample undergoes a different optical path length as the light from the reference mirror, however interference is only observed when the difference in optical path length lies within the coherence length of the light source. This interference pattern can be used to reconstruct a one-dimensional reflectance profile of the sample along the direction of the beam, which is called *axial scan* (A-scan). The imaging speed of an OCT system is typically characterized in terms of the A-scan rate as A-scans per second.

In the following, we give a short overview of the differences of the three OCT variants as far as they are relevant to the contents of this thesis, and refer the interested reader to the excellent compendium edited by Drexler and Fujimoto [96] for further details. The three variants differ in how the reflectance profile that forms an A-scan is achieved and represent the three major advances of OCT imaging technology:

Time Domain OCT (TD OCT) relies on physically scanning the reference mirror, thereby modifying the zero-delay position. The detector, a single photodetector in this case, only measures interference from the the signal reflected at the tissue depth corresponding to the reference path length. Therefore, the movement of the reference mirror over time creates a reflectance profile along the depth of the beam. TD OCT was the first generation of OCT devices and was mainly constrained by the mechanical motion of the reference mirror as well as the low sensitivity of the detection. This limits TD OCT to relatively slow A-scan rates of up to 2kHz [203].

Spectral/Fourier Domain OCT (SD OCT) changed the detection scheme by replacing the photodetector with a spectrometer together with a high speed line camera. The diffraction grating of the spectrometer decomposes the spectral components of the combined light coming from sample and reference arm. The line camera detects these spectral components. This allows the simultaneous detection of the complete axial reflectance profile without motion of the reference mirror. As the name suggests, the line camera detects the signal in the spectral domain. Thus the measured signal has to be transformed using a Fourier transformation to obtain the actual reflectivity profile. SD OCT was first suggested in 1995, however widespread interest only sparked in 2003 when it was shown that its simultaneous detection of all echoes can lead to sensitivity increases of 50-100 times compared to TD OCT [45]. This increase in sensitivity allows for an equivalent increase in measurement speed and current SD OCT systems reach high A-scan rates of 147kHz [234] and up to 312kHz with $8.7\mu m$ resolution in tissue [165], however these imaging speeds generally come at a cost of axial resolution and sensitivity due to limitations of the line scan cameras.

Swept Source/Fourier Domain OCT (SS OCT) also measures the axial reflectance profile in the spectral domain, however in this detection method a narrow-bandwidth, frequency-swept light source is used. Instead of a spectrometer splitting the spectral composition of a broadband light source, a photodetector measures the signal corresponding to the frequency the laser is tuned to at a time point. The frequency of the light source is swept through the spectral range and samples of the photodetector signal are acquired during this sweep with a high resolution analog/digital converter. This set of samples is equivalent to the set of spectral domain measurements acquired from the line camera in SD OCT and is again transformed

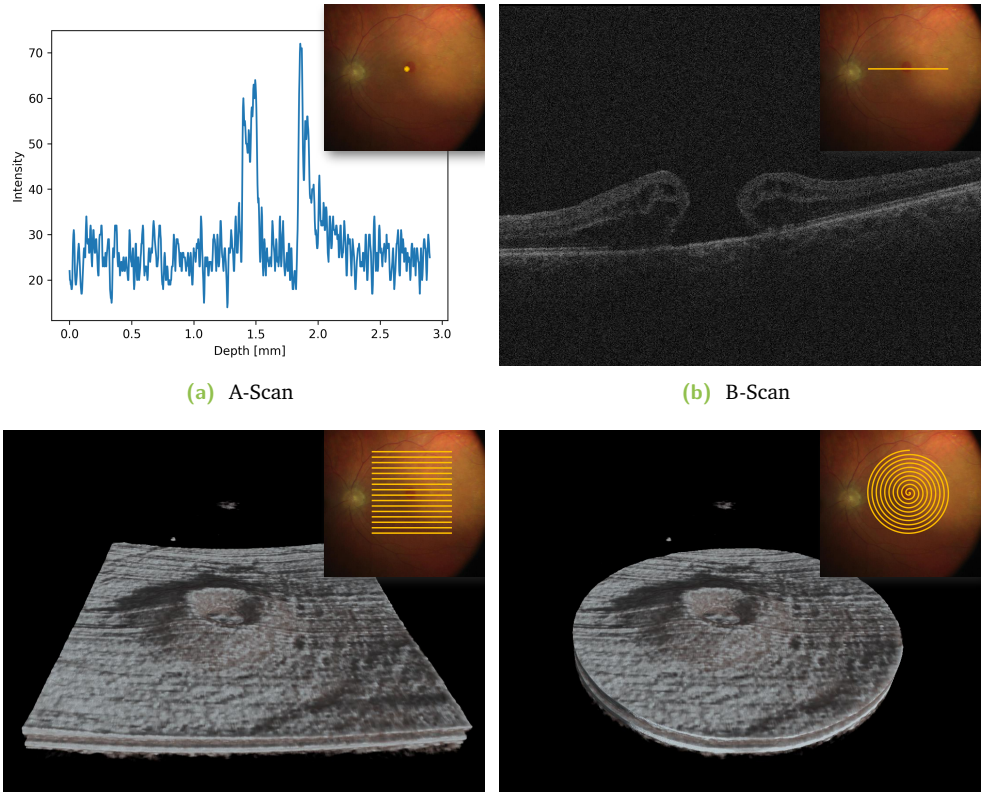
with a Fourier transformation to obtain the axial reflectance profile. The A-scan rate of SS OCT is mostly limited by the sweep rate of the light source, and advances in the domain of Fourier-domain mode locking (FDML) lasers and vertical cavity surface-emitting lasers (VCSEL) have brought sweep rates (and thus A-scan rates) up to multiple MHz[20, 111]. SS OCT light sources generally have a much longer coherence length which increases the possible imaging range to several centimeters and enables whole eye imaging, including the anterior and posterior region in one scan [73]. VCSEL light sources offer additional flexibility as the frequency sweep is tunable via a microelectromechanical system (MEMS). This enables dynamic tuning of the depth imaging range, resolution and axial scan rate [45] and thus provides a great amount of flexibility for different applications within the same device [20].

OCT Image Compounding

While some applications like instrument-integrated OCT [33] purely rely on the information provided by single axial scans, generally the imaging systems combine several A-scans together to form images, as illustrated in Figure 2.9. This is achieved by deflecting the sample beam with an arrangement of two mirrors for x and y axis deflection to scan the sampling beam across the tissue. Scanning the beam along a line provides 2D images which are called B-scans analogous to ultrasound B-mode imaging. Further combination of several parallel B-scans can be achieved by raster scanning the beam to create 3D volumes, sometimes called C-scans. The introduction of ultrahigh speed SS OCT systems has shown that conventional raster scanning techniques are limited with respect to acquisition dead times as well as sampling uniformity, which has lead to the development of spiral scanning modes [25]. In this scanning mode, the sampling beam is deflected in a spiral pattern, resulting in a cylindrical acquisition volume with uniform sampling. The increased scan efficiency and reduced mechanical stress on the deflecting mirrors make this an attractive scanning mode for continuous 4D volumetric acquisition. Spiral scanning however requires an additional resampling step to transform the A-scans from a cylindrical coordinate system to a standard Cartesian grid.

OCT systems generally are limited by a certain A-scan rate, which in practice results in a trade-off between imaging area, update rate and over/undersampling: when a higher update rate is desired, this can only provided by either only scanning a limited area of the tissue or conversely undersampling the tissue, distributing the A-scans further apart than their transverse resolution. Conversely, if high image quality is required, multiple A-scans of the same tissue location can be combined to an averaged measurement at the cost of reduced imaging speed. This approach is often used in diagnostic scenarios where real-time feedback is not required.

The scanning mirrors are controlled electronically which allows dynamic control of the scanning pattern, allowing an operator to choose between scanning modes intended for different use cases. For example, the RESCAN 700 (Carl Zeiss Meditec, Jena) intraoperative SD OCT provides three continuous scanning modes for retinal surgery: a single B-scan consisting of 512 A-scans updating at $\sim 50Hz$, a cross pattern of two orthogonal B-scans at $\sim 25Hz$ and five parallel B-scans at $\sim 10Hz$. In addition, a raster scan of 512×128 A-scans can be initiated which requires $\sim 1-2s$ acquisition time and is intended mostly for documentation purposes.



(a) A-Scan (b) B-Scan

(c) Volume raster scan (d) Volume spiral scan*

Fig. 2.9. OCT Beam scanning modes to create different images: A-scans provide reflectance along a depth profile for a single point on the retina **(a)**. B-scans form a 2D cross-section into the retina **(b)**). Multiple parallel B-Scans can be combined to perform volume raster scanning **(c)**. Alternatively, a spiral scanning pattern can be used to obtain a cylindrical imaging volume **(d)**. *The 3D rendering of the spiral scan has been simulated from raster scan data to approximate the spiral scan field of view.

2.2.2 Intraoperative Optical Coherence Tomography (iOCT)

The general potential of providing high-resolution in-vivo imaging for intraoperative image guidance has been realized early on and has led to the commercialization of microscope-integrated OCT in 2014/2015 [26]. In these systems, the deflecting mirrors for transverse beam scanning are integrated into the surgical microscope and the sampling beam is fed into the objective lens of the microscope for seamless imaging of ocular structures during the surgical procedure. The real-time feedback for intraoperative maneuvers enables individualized patient care, and its potential to support novel procedures has been recognized already shortly after the commercial introduction of such devices [53].

The current generation of commercially available devices use SD OCT engines with A-scan rates of 27-35 kHz, which enables live B-scan imaging at real-time update rates (15-60Hz, depending on the scanning pattern and resolution). Volumetric acquisition at high resolution is only provided in a stop-and-shoot manner as higher resolution volumes (typical sizes are 200×200 or 512×128 A-scans) require several seconds. Real-time volumetric feedback is not provided by the current commercial systems, as the A-scan rates would limit this to either low update rates or low scanning resolution. It is however expected that the trend in diagnostic

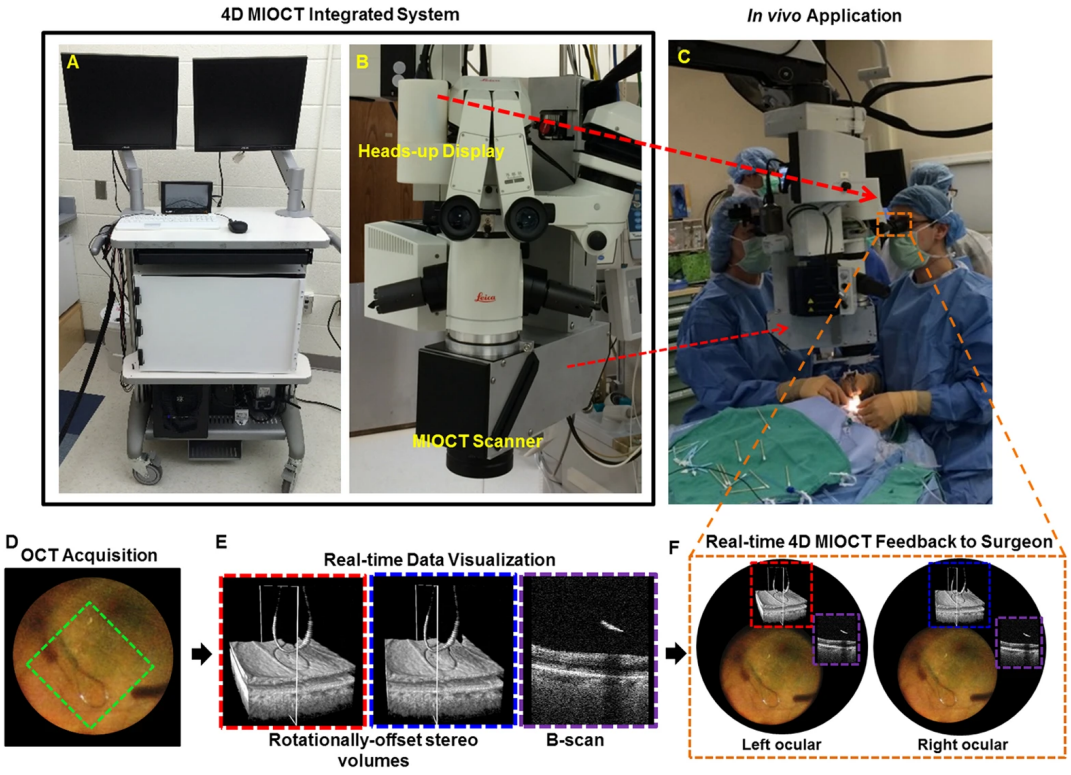
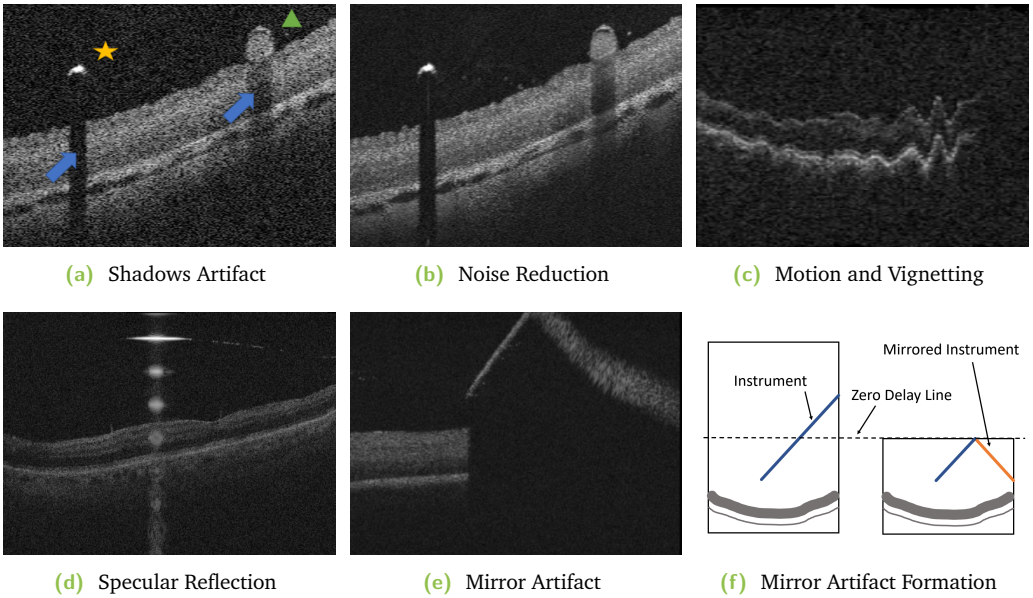


Image source: [24]

Fig. 2.10. Overview of the 4D microscope-integrated OCT system at Duke University [24]: Their 4D OCT system is housed in a portable cart **(A)** and the imaging laser is directly coupled into the microscope **(B)**. They demonstrated their system in human surgery, performing live 4D volumetric imaging and stereoscopic live visualization in the ocular **(D-F)**.

OCT systems to switch towards SS OCT engines will also apply to intraoperative systems, thus enabling much higher imaging speeds. Prototype systems for high-speed intraoperative imaging have already been presented by several research groups: Kang et al. at Johns Hopkins University presented a 128kHz SD OCT system tailored towards microvascular anastomosis [235], however their system was not integrated with an operating microscope. Probst et al. at University of Lübeck reported a microscope-integrated OCT system capable of 4D imaging at 7Hz volume rate [166] with a volume resolution of 300×80×512 using a 210kHz SD OCT system, however the low imaging sensitivity severely limited the potential applications. Later, Li et al. used a 50kHz SS OCT for 4D image guidance of glaucoma surgery [124]. At the time of writing, the currently most mature research iOCT system is presented by the group at Duke University. Carrasco-Zevallos et al. developed a 100kHz SS OCT system integrated into a surgical microscope capable of real-time 4D imaging [27] which they later evaluated in *in vivo* human study [24] for anterior and posterior tasks, performing imaging at a resolution of 300×100×595 voxels at 3.33Hz update rate. An overview of their system is depicted in Figure 2.10. The same group recently presented a novel system [212] with a 400kHz A-scan rate which is intended to be translated to human ophthalmic surgery after an initial evaluation phase with mock surgeries.

The display of intraoperative imaging information to the surgeon poses a novel challenge. Visualization of the data can be provided on an external screen which provides great flexibility



(a) Shadows Artifact (b) Noise Reduction (c) Motion and Vignetting

(d) Specular Reflection (e) Mirror Artifact (f) Mirror Artifact Formation

Fig. 2.11. Common imaging artifacts of intraoperative OCT: **(a)** B-scan depicting shadowing artifacts (arrows) of instrument ⋆ and blood vessel △. **(b)** Averaging (10x) increases SNR at the cost of decreased imaging speed. **(c)** Axial motion during acquisition leads to distorted surfaces. Vignetting (right) causes signal loss in some A-scans. **(d)** Specular reflection at a fluid-gas interface (faint line visible in the top right) creates repetitive columns of high signal. **(e,f)** Mirror artifact causes structures in the negative light delay space to be folded into the image as if mirrored at the zero-delay line.

and the option for stereoscopic display if an appropriate screen is used. This however has the downside that the surgeon needs to switch between the ocular direct view of the microscope and the view of the external display, leading to shared attention and reduced surgical stability unless the external display also can provide the microscopic view. An alternative is the integration of semi-transparent displays inside the microscope, which can then overlay the information directly inside the microscopic view. The Lumera 700 with integrated RESCAN 700 (Carl Zeiss Meditec, Jena) uses a monocular display to this end and Carrasco-Zevallos et al. use a stereoscopic variant [24]. Draelos et al. have shown the concept of using an immersive VR system (HTC Vive) [42] to show real-time 4D OCT combined with a video passthrough of the headset to retain real-world spatial awareness of the user in anterior surgery. Their system does not use a microscope-integrated OCT so no microscopic image is combined.

Imaging Artifacts

OCT imaging exhibits several typical artifacts that complicate consistent visualization of 3D imaging data and iOCT introduces additional challenges due to the dynamic nature of intraoperative scenario. Clinical analysis of various artifacts in diagnostic OCT has been performed in various studies [6, 80, 195, 242] to assess their impact in a diagnostic scenario. Here we provide an overview of the most common artifacts inherent to (volumetric) intraoperative OCT to illustrate the challenges involved in providing real-time visual feedback.

Speckle Noise OCT is subject to speckle noise which can be best modeled by a Gamma Distribution [108]. While the noise texture potentially contains useful information about tissue micro-structure [108, 183], in rendering it presents a challenge mostly for classification

when an intensity- or gradient-based transfer function is applied. Denoising solutions have been proposed both by hardware modifications [125] and image processing [1, 7, 122, 182]. One effective way to reduce noise is averaging of several co-located B-scans as seen in Figure 2.11b, however this results in reduced effective B-scan rate. Nonetheless, only very few can be efficiently applied within the tight time constraints of an intraoperative scenario so commonly only simple processing like Gaussian, median or bilateral filtering are applied.

Shadows Given its directional nature, OCT cannot always reconstruct tissue when the imaging beam hits highly absorbing or reflecting structures. This results in directional shadows below these opaque structures, which are often blood vessels (causing diminished signal) or metal surgical tools (causing complete signal loss) as can be seen in Figure 2.11a. Instrument shadows are especially critical as they can hide important tool-tissue interactions. The use of materials that are transparent to near-infrared light to build instruments can mitigate this problem [52], however these have yet to be evaluated in broader clinical studies and are therefore not easily available. Vignetting, as seen in the right side if Figure Figure 2.11c can also be interpreted as a form of shadowing caused by the iris, leading to total signal loss for larger areas in the region of interest.

Specular Reflections Specular reflections (see Figure 2.11d), analogous to specular highlights visible in cameras, are caused when the camera is near the reflection angle with respect to the light source. In OCT, as the light source and detector are co-located, this happens when a highly reflective surface (such as a gas-fluid interface) is nearly perpendicular to the imaging direction.

Mirror Artifact The mirror artifact, depicted in Figure 2.11e, is a common artifact of both Fourier domain OCT technologies that arises from the image signal reconstruction through a Fourier analysis. The A-scan reconstruction cannot separate backscattered light of positive and negative light delay and thus a reconstructed A-scan is always a superposition of both echos around the zero delay line (see Figure 2.11f). When imaging retinal structures, this is not necessarily problematic because the negative delay part contains only "empty" vitreous and the retinal structure is usually kept within one part of the range. Problems arise when structures cross the zero-delay line as one half will then appear folded, or mirrored, into the image. This can be confusing when approaching the retinal structure with a surgical instrument, as it first shows mirrored "upside down" while it is in the negative delay space until it is close enough to enter the positive delay space. Mirror artifacts have also been shown to have a higher occurrence in highly myopic eyes [91] even in the absence of instrument.

Motion Artifact Raster scanning to acquire volumetric OCT introduces a *fast* axis (along the B-scans) and a *slow* axis perpendicular to the B-scans. For the acquisition of one B-Scan, eye motion is negligible, however motion between subsequent B-Scans results in wavy patterns along the *slow* axis of a volume if the motion is fast compared to the volumetric scanning time. Figure 2.11c depicts this effect. During surgery, the patient's eye is usually immobilized to prevent involuntary eye movement (microsaccades). However, the patient's other vital functions like cardiac cycle or breathing [195] as well as the surgeon manipulating surgical instruments in the trocars introduce movement in both axial and lateral directions.

Uneven Contrast Limitations in the data acquisition and reconstruction lead to uneven contrast within the image. This artifact is mostly a problem in SD OCT, where there is a strong sensitivity falloff with respect to the distance to the zero-delay line. For swept-source systems, this depth sensitivity falloff depends on the light source used and is generally less pronounced than in SD OCT, especially for VCSEL lasers which create a mostly stable sensitivity across the whole depth range.

External Image Quality Factors

Retinal OCT, especially in the intraoperative case, suffers from unstable image quality and intermittently low SNR. This is caused by the highly dynamic operating environment where the optical pathway changes due to manipulations of the surgeon which can significantly affect image quality. We have found that the following potential factors can impact overall image quality during a surgery, sometimes dramatically:

Focus/Defocus The focus of the microscope optics is determined not only by the optics internal to the microscope, but also by the optical components of the patient's eye, mainly the cornea and lens. Since the eye is not a static part of the optical pathway, this can change intermittently, due to various factors: Moving the eye mechanically is a common way for the surgeon to control the visualized retinal region, but it might also happen involuntarily during manipulation through lateral forces on the trocars. Additionally, the optical properties of the eye itself can change, for example by varying intraocular pressure after vitrectomy (replacement of the vitreous humor with liquid or air) or after application of transparent gel that prevents drying-out of the cornea. Most microscope-integrated OCT engines have a separate focus for the OCT beam to compensate for discrepancies between focus of the overall microscope optics and the OCT image region. However, at least in present-day commercially available devices, this focus has to be managed manually. Because neither nurses nor doctors can constantly monitor optimal focus, this often leads to tissue being partially or completely unfocused.

Polarization To optimize the coherence pattern, the sample and reference arm of the OCT need to be equally polarized (see section 2.2.1) for optimal sensitivity. In practice the polarization can become misaligned during a surgery due to mechanical changes of the operating microscope deforming the optical fibers. Motorized polarization paddles are used to re-align sample and reference arm, however these are not adjusted contiuously which can lead to intermittent suboptimal sensitivity and decreased signal-to-noise ratio.

Distortion Geometric distortions make it hard to relate measurements in OCT to real-world structures. One big source of distortions is the fan-beam geometry of the OCT scan [224]. While it is possible to calibrate and correct these distortions in diagnostic devices [32, 163, 224], optical pathways are more complex in retinal iOCT as they include not only the variable microscope zoom and focus optics but also the anterior segment of the patient with potential motion of the lens from rotation of the eye. Thus iOCT data is commonly displayed uncorrected which leads to noticeable distortions for large scan sizes where the opening angle of the fan beam is large.

2.3 Volume Rendering

Rendering of 3D volumetric data is an important technique for many medical applications. A straightforward way to use the utilize the mesh-based rendering pipeline of modern GPUs is to extract surface meshes using segmentation algorithms or simple thresholding. While this is a convenient way to show prominent surfaces in the data set, it quickly becomes limiting when multiple surfaces are involved or less defined, volumetric structures are involved. The alternative to surface extraction is rendering the volumetric data directly without an intermediate mesh representation. There are different rendering algorithms to accomplish this, like cell projection, shear-warp projection, volume splatting, volume raymarching or monte carlo integration. They share the common notion that rendering of participating media can be achieved by approximating the volume rendering integral using an emission-absorption model [139]. In this optical model, each position $\mathbf{x} \in \mathbb{R}^3$ in the volume is associated with an emissive component $c(\mathbf{x}) \in \mathbb{R}^3$ and an absorption coefficient $\kappa(\mathbf{x}) \in \mathbb{R}$. Computing the radiant energy C incident on a virtual camera's pixel is then formulated as evaluating an integral along the camera ray $\mathbf{x}(t)$:

$$C = \int_0^\infty c\left(\mathbf{x}(t)\right) \cdot e^{-\tau(t)} dt \tag{2.1}$$

where $\tau(d) = \int_0^d \kappa(\mathbf{x}(\hat{t}))d\hat{t}$ can be interpreted as a visibility term of a position, accumulating the amount of absorption between the camera and the position $\mathbf{x}(d)$ along the ray.

2.3.1 Direct Volume Rendering

Direct volume rendering (DVR), commonly implemented with raymarching, is an efficient way to approximate the above integral via numeric integration. The discrete approximation of Equation 2.1 can be reformulated as a compositing problem. The volume raymarching algorithm [140] combines equidistant samples at $t_i = i\Delta t$ from front to back along the ray using the well-known alpha blending equation. In this case, the emitted color C_i and opacity A_i at sample point i is defined as

$$C_i = c(i\Delta t)\Delta t, \tag{2.2}$$
$$A_i = 1 - e^{-\kappa(i\Delta t)}\Delta t. \tag{2.3}$$

Iterative evaluation can be achieved from front to back with alpha blending as

$$C'_i = C'_{i-1} + \left(1 - A'_{i-1}\right) C_i, \tag{2.4}$$
$$A'_i = A'_{i-1} + \left(1 - A'_{i-1}\right) A_i \tag{2.5}$$

with starting conditions $C'_0 = 0, A_0 = 0$.

The final color for a pixel is then determined as C_N, A_N by accumulating all N samples along the viewing ray that fall within the volume bounds.

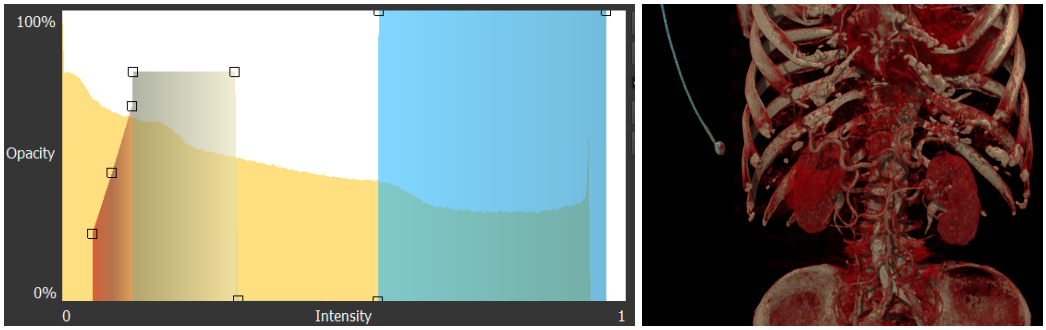


Fig. 2.12. Example of an intensity-based transfer function editor and resulting rendering. The user can edit the mapping of intensity to color/opacity by modifying the keypoints (small squares) or adding/moving the whole geometric primitives. For example, the red trapezoid on the left covers a data range that corresponds to soft tissues while higher intensities correspond to more dense, bony structures (gray) and eventually metallic structures (blue).

2.3.2 Transfer Functions

The mapping from the raw volume data at each sample point to an associated emissive color and absorption coefficient is usually controlled by a transfer function. A user interface allows users to modify this mapping, often through the use of geometric primitives with associated colors to control the appearance of different structures in the volume. An example is shown in Figure 2.12. The input of a transfer function is a set of features extracted from the volume. In the simplest case, the (normalized) intensity of the volume is used, however intensity alone is often ambiguous and not always sufficient to discriminate tissue types. A vast set of additional features such as gradient magnitude, curvature, occlusion, vesselness, semantic labels, additional modalities, geometric features and other derived features have been investigated as a means to discriminate different tissues in various applications [129]. Transfer functions can be decomposed into a classification step (e.g. identifying specific structures from the input features) and material mapping step (e.g. associating specific visual attributes like color and absorption with each of the structures). In a common transfer function user interface, the classification step is represented by specifying ranges in the input feature space with geometric primitives, often so each primitive covers one semantic class. The material mapping is then performed by associating color and opacity with each primitive to specify the visual appearance of each structure.

2.3.3 Advanced DVR Effects

The basic emission-absorption model is a rather simplified illumination model and only approximates light contributions emitted by the sample voxels, ignoring any scattering effects that occur in real participating media. Move advanced illumination models have been used to produce higher-quality renderings by considering single and multiple scattering. In medical image analysis, producing realistic renderings not only helps clinicians better correlate the visual information to real patient anatomy but also improve spatial understanding as shading and shadowing are important cues. Gradient-based shading of volumetric surfaces has been employed as early as [120]. The use of Phong shading for implicit surfaces is popular, however as surface shading techniques they do not have a strong justification for use in soft participating

media. In a physically-based formulation of volumetric light transport [162], the total radiance L_s from a voxel at position p in direction ω can be expressed in relation to the emission L_e, absorption σ_s and phase function p as

$$L_s(p,\omega) = L_e(p,\omega) + \sigma_s(p,\omega) \int_{\mathcal{S}^2} p(p,\omega_i,\omega) L_i(p,\omega_i) d\omega_i. \tag{2.6}$$

$L_e(p,\omega)$ is the amount of light emitted from the object itself towards ω. The phase function $p(\omega_i,\omega_o)$ characterizes the scattering behavior and determines how much of the incoming light L_i is scattered from an incoming light direction ω_i towards an output direction ω_o. The integral evaluates this for all directions $\omega_i \in \mathcal{S}^2$ in the sphere of directions to obtain the total out-scattered light in direction ω. The absorption function σ_s determines how much of the in-scattered light is lost due to out-scattering or absorption.

Advanced volume rendering algorithms that implement single or multiple scattering provide shadows and enhance realism of the view. Numerous efficient real-time algorithms to approximate volumetric shadows from single and multiple light sources have been developed over the years [77, 89, 101, 102, 174, 198, 199]. More realistic global illumination approaches approximate illumination probabilistically using Monte-Carlo integration[113] to create highly realistic images, however these are not easily computed in real-time.

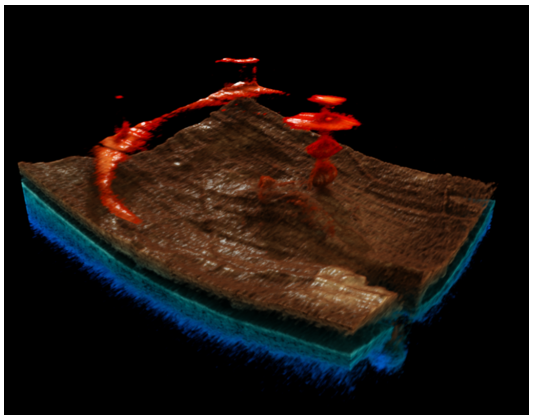
2.3.4 Challenges of 3D iOCT Visualization

Volume rendering of OCT data is particularly demanding due to the multitude of potential artifacts that are inherent in the modality. This section discusses how the OCT artifacts described in section 2.2.2 affect volume rendering of 3D and 4D OCT scans.

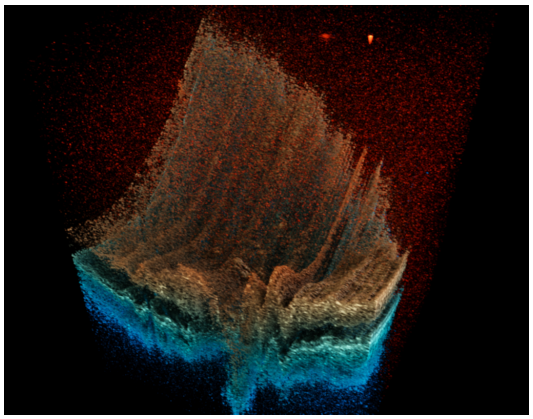
Structural inconsistencies Artifacts that are familiar to a surgeon in a 2D B-scan manifest in a 3D volume as pseudo structures which are harder to interpret. For example, the specular artifact seen in Figure 2.13b creates distracting vertical columns of high intensity on the retinal tissue. While this is relatively easy to identify by experienced users, it still creates a large structure which can occlude important information. The mirror artifact does not directly create non-existent structures (c.f. Figure 2.11e, but rather shows them upside down at a wrong position and with inverted axial movement direction. An instrument at a safe distance from the retina but inside of the negative delay region of the OCT can appear to be close or inside tissue. Movement of tissue or instrument during acquisition creates displacement between adjacent B-scans. Movement in axial direction usually manifests as "wavy" surfaces in a volumetric scan by axial displacement of subsequent B-scans. Lateral movement creates visuals akin to a rolling shutter effect in digital photography [34] and often affects instruments which appear stretched or discontinuous.

Missing information Tissue can also be hidden by the shadow artifact which is especially relevant intraoperatively, as exemplified in Figure 2.13a. Signal from tissue beneath a metal instrument is completely blocked, leading to imaging blind spots in the retinal surface close to the location of intended manipulation. Vignetting of the OCT imaging area by anterior structures can cause signal loss for larger parts of the volume as seen in Figure 2.13c.

(a) Needle touching retinal surface from right to left, showing mirror artifact (inverted right side of the needle) and instrument shadowing (gap below needle).



(b) Specular reflection (central column) at gas-vitreous interface.



(c) Axial motion artifact (wavy surface appearance) and vignetting (missing top right corner of the volume).

Fig. 2.13. Common iOCT artifacts as they manifest in volume rendering. Images are rendered with a Monte Carlo renderer based on [113]. Colorization based on axial position as further described in section 8.

Intensity variations The OCT imaging modality only measures backscattering and reflectance, thus similar OCT intensities do not necessarily correspond to structural similarities. Intensity-based transfer function classification therefore often cannot discriminate consistently between structures like different layers. The problem is compounded by the combined effects of uneven imaging sensitivity across the axial distance, partial shadowing through for example blood vessels and dynamically changing contrast through changes in the optical system. For this reason, transfer functions often have to be adapted for individual volumes and even in response to dynamic changes during an operation. The varying amount of speckle noise also makes usage of the gradient as an additional feature unreliable not only as a way of classification, but also for shading methods that rely on a surface normal extracted from the gradient.

4D OCT

Many of these challenges can be resolved in the diagnostic scenario where enough computational power is available and little time constraints are imposed. However, most of these solutions are not applicable for intraoperative imaging. An interruption of the surgical manipulation for imaging and review should be kept to a few seconds while feedback during critical maneuvers should ideally be instantaneous and without requiring interaction with the visualization. With research-grade 4D-capable OCT systems, some problems like movement artifacts and mirror artifacts are less pronounced due to increased imaging speed and depth. However, ultrafast SS OCT imposes additional challenges related to data processing. High performance systems as for example reported by Kolb et al. [111] scan at a data rate of 1.58 GVoxels/s, a bandwidth of 6GB/s. While modern GPUs can still render this data directly in real-time with basic emission/absorption direct volume rendering, these data rates severely constrain any additional image processing or more advanced rendering. Careful pipeline design and efficient GPU processing is required to process the imaging data in real time without introducing latency between acquisition and display.

Part II

Towards an Integrated Robotic Retinal Microsurgery Environment

The introduction of novel robotic solutions will allow retinal physicians to perform surgeries with less complications and enable novel interventions that were previously infeasible because of limiting human factors. The superior precision and stability offered by retinal microsurgery robots hold the promise of precise manipulation of retinal tissue and targeted delivery of therapeutic vectors. Yet to fully leverage the micrometer precision offered by such systems and move towards more autonomous systems, the loop needs to be closed by combining the robotics with sensing technologies of matching resolution. Fully autonomous surgical systems may still be in the rather distant future, therefore a human in the loop will still be required for some time to operate the robotic systems. Even though a rising level of robotic automation might reduce the amount of manual interaction such an operator will have with the system itself, humans will always need to rely on good visualizations in order to ensure the system performs its job adequately.

The stereoscopic surgical microscope traditionally used to support retinal surgery can only provide an enface view of the retina. Perspective and complex non-linear optics impede distance judgements between instrument and tissue for human observers and computer-vision based approaches [230] alike. Intraoperative OCT can provide complementary depth-resolved information and is therefore expected to play a major role in this robotic transformation [210].There are only very few examples of integrative advances between robotic and iOCT technology for retinal microsurgery. They have for the most part been been focused on instrument-integrated OCT [210] to augment sensing at the instrument tip. Zhou et al. have investigated the integration of OCT into the robotic control loop from a technical side, considering registration of robot controlled instrument and OCT coordinate systems [238] and needle localization for OCT-guided subretinal injection [239] as necessary components for closed-loop robot control. However, to create an optimal environment for iOCT-guided robotic retinal surgery, it is necessary to integrate all components in a holistic approach, combining the robotic control not only with image-based feedback but also with a visualization and user interface tailored to this specific need. In such a system, visual feedback for the operator plays a crucial role to avoid misinterpretation and provide sufficient information to make split-second decisions during interventions.

In the following sections, we present a first concept for such a fully integrated system. This proposed system combines live 2D/3D OCT imaging with a teleoperated robot control. Specialized visualizations show not only the current imaging data but also integrate path planning information of the robot control software to allow for real-time monitoring in settings with more robotic autonomy.

We will first describe a modified workflow for image-guided robotic sub-retinal injection which was developed as a result of discussions with clinical experts in section 3. Section 4 then describes the visual guidance prototypes that we developed for supporting such a workflow, considering first a traditional system based on a stereo screen (section 4) and secondly the use of an immersive VR headset (section 5) to create an information-rich environment for the surgeon. In section 6 we will then translate the learnings from these visual prototypes into component requirements for effective visual guidance in a fully integrated visual guidance system. These will form the central motivation for the methodology presented in Part III.

3 Image Guided Robotic Surgery Workflow

Contents

In this section, we will describe a possible surgical workflow for robotic image-guided surgery as developed in cooperation with expert retinal surgeons as well as robotic domain experts. A bespoke workflow was defined that optimally leverages the strengths of intraoperative imaging in conjunction with a precise robotic manipulator. This vision of how robotic systems can be applied in practice forms the motivation and basic framework for the design exploration studies in the following sections.

Robotic autonomy can be characterized on a continuum from fully manual to fully automated procedures[13, 231]. Yip et al. [231] divide this spectrum into four categories:

1. *Direct Control*: The surgeon exclusively controls the robotic effector, for example in normal teleoperated systems.
2. *Shared Control*: The robotic system can intervene in the motion, for example to stabilize motion or restrict positioning to safety margins.
3. *Supervised Autonomy*: The robot executes tasks by itself while a human supervisor remains responsible and in charge of decision making.
4. *Full Autonomy*: The robot fully controls all aspects and independently makes all decisions without external supervision.

We designed our integrated robotic environment and workflow for a teleoperated robot with the goal to create a visually assisted workflow that focuses on *shared control* with the opportunity to go towards *supervised autonomy*. The system targets precise subretinal injections as described in section 2.1 as the main use case.

Perioperative Planning

A central step of the workflow is a perioperative planning stage to perform thorough analysis of the injection site and precise planning of the robotic motion. This is consistent with the current standard operating workflow of retinal procedures, were extensive preoperative planning is rarely performed due to the complex changes in retinal morphology induced by the vitrectomy step. Pre-operative planning based on diagnostic OCT would therefore require an additional registration step between the pre-operative and intra-operative imaging. This registration step would add an additional source of error which we avoid to ensure micrometer precision of the

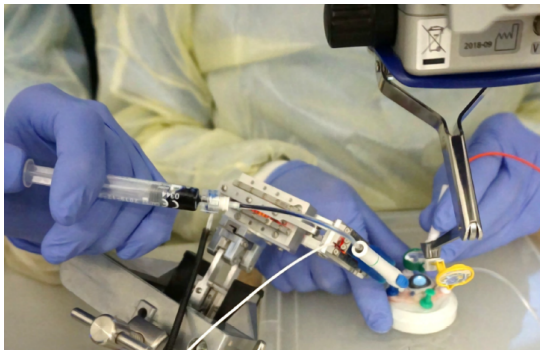


Image source: [240] ©2019 IEEE.

(a) iRAM!S robotic platform

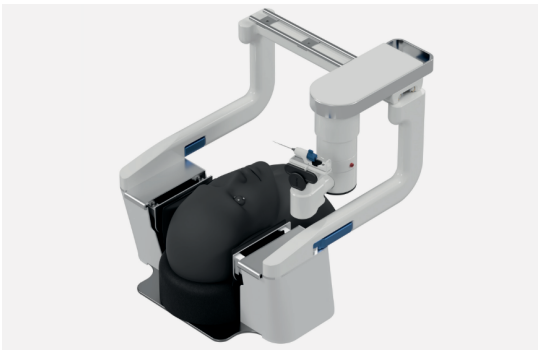


Image source: [222] ©2019 M. Weisser

(b) Design concept for the RASIS microsurgery system.

Fig. 3.1. The robotic platform used for design and experimentation of our robot-integrated visual workflow. iRAM!S is the teleoperated robot and RASIS integrates this robot into a bed-mounted system for integration into the surgical environment.

overall injection. The perioperative planning step in our workflow is enabled by OCT as a high-resolution intraoperative imaging modality and ensures that ad-hoc planning is performed on the latest morphology after vitrectomy. The goal of the planning step is to provide the intended injection location as well as potential risk areas to the robot control software module. After the planning step, the execution of the robotic maneuvers is performed under visual guidance to provide confirmation of the precise instrument placement and planned direction of movement.

Robotic platform

As a platform for our experiments and design considerations, we used the teleoperated iRAM!S robot (figure 3.1a) with 5DoF movement and a digitally controllable remote center of motion (RCM) [153, 240]. The master system features a conventional joystick and the control mode can be switched between unconstrained motion for alignment and insertion into the trocar and RCM restricted motion for precise control inside the eye while avoiding motion lateral to the trocar. An external touch screen computer is used to modify additional parameters of the robot such as maximum translational and rotational velocity and RCM modes. The slave system receives control commands via a network interface and converts these inputs to the control signals for the piezoelectronic linear actuators of the robotic system. The RASIS microsurgical system design (figure 3.1b) allows for further degrees of freedom for rough alignment of the robot with the patient eye [222], however this design has not been fully realized yet.

3.1 Surgical Environment

The surgical environment of our proposed image-guided robotic surgery system consists of several additional components compared to the standard setup. The major components are depicted in a lab setup in Figure 3.2. For a full surgery setup, they include:

- Anesthesiology equipment

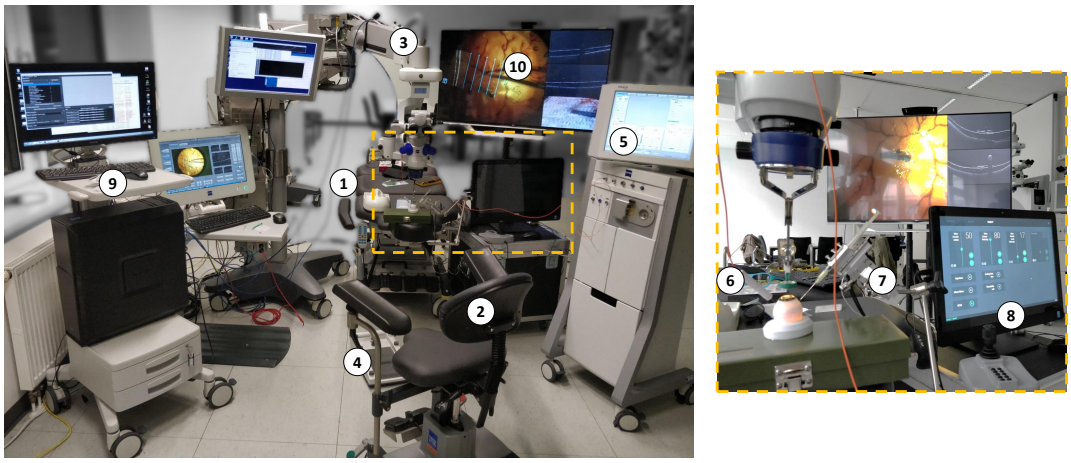


Fig. 3.2. Equipment involved in a lab setup for an image-guided robotic surgery.

- Patient bed (1)
- Surgeon chair with armrests (2)
- Stereo microscope with integrated OCT (3) and control foot pedal (4)
- Vitrectomy equipment (5) and control foot pedal (hidden in figure)
- Endoilluminator light source (6)
- Robot slave system including motor drivers (7)
- Robot master system with joystick controller (8)
- Advanced visualization system including PC control cart (9) connected stereoscopic visualization screen (10)

The main novel additions to the conventional surgical setup described in section 2.1 are the robotic components (7),(8) and the visualization setup (9),(10). The visualization PC collects imaging data from the stereo microscope cameras as well as real-time OCT imaging. The visualization application combines intraoperative and planning information to provide a stereoscopic visualization on the surgeon screen (10). The master robot system is connected via a network interface to the visualization system to exchange robot status information and receive data from the intraoperative imaging systems relevant to the robotic control.

Hardware Platform

To realize a prototype interactive visualization system, we use a Lumera 700 with integrated OCT (RESCAN 700) as the microscope base. Two Grasshopper GS 3 (Teledyne FLIR LLC, Wilsonville, USA) USB cameras (2048×2048 @30Hz) were mounted on the stereo microscope to allow for digitization of the microscope view. The pair of cameras is connected to the visualization system PC (Intel i7-8900K, 64GB RAM, NVidia GTX 980Ti) which runs the visualization applications and outputs images to a passive stereo screen with a total resolution of 3840×2160 px. The screen uses horizontal interlacing of left and right view, effectively halving the vertical resolution to 1080 lines when in stereoscopic mode. The OCT data from the RESCAN 700 iOCT system is transmitted via a 1000 Mbit/s ethernet connection through a special firmware version customized for research purposes.

3.2 Modified Surgical Workflow

Compared to a conventional injection procedure, we replace the manual injection step with a two-step robotic procedure divided into perioperative planning and execution. The surgical workflow is thus roughly divided into four main phases:

1. Surgical Preparation The preparation phase includes the traditional surgical setup of patient sedation and preparation of the sterile environment. The eye clamps are placed and the traditional three-port trocar setup is prepared. Vitrectomy is performed to replace the vitreous gel with balanced saline solution (BSS). This is necessary because instrument movement within the vitreous can cause traction on the retinal tissue with adverse side effects like retinal tearing or detachment. After this initial setup, the manually operated endoilluminator used during this phase is replaced with a chandelier light source fixed in the trocar. A fixed light source is preferable in robotic interventions to prevent introducing external movement from a manually operated endoilluminator. The preparation phase concludes with the mounting and introduction of the robotic manipulator into the surgical area and the instrument trocar. After these steps, the eye is fully stabilized as the only motion in the surgical area is from patient-intrinsic movement (breathing and heart rate). Direct interaction with the surgical area is no longer required with the exception of occasional lubrication of the corneal tissue (usually performed by the assistant) to avoid it drying out and preserve optimal viewing conditions.

2. Periprocedural Planning The planning stage begins with the acquisition of a high resolution OCT cube of the retinal area where the injection will be performed. This OCT cube is processed by the visualization system and presented to the surgeon on the stereoscopic display. The visualization application provides a planning interface that can be controlled with a normal computer mouse. In this planning interface, the surgeon is able to define the optimal needle insertion position and depth within the retinal tissue. The details of this planning environment is described in detail in section 4.1. The intervention plan and imaging information is then communicated to the master system of the robot, which then uses this information during the execution phase to optimize the robot motion.

3. Robotic Execution In this critical phase, the robotic intervention is performed. The visualization is changed to an optimized view for intraoperative guidance further explained in section 4.2. In this phase, the kind of manual control for the surgeon differs by whether the robotic system is operating in *shared control* mode or under *supervised autonomy*. In *shared control*, the robotic system merely monitors the instrument tip position and ensures that no involuntary deviation from the planned injection trajectory occur while the surgeon controls the motion with the joystick controller. In the *supervised autonomy* case, the surgeon only confirms the pre-planned robotic motion by holding a button to advance the system while the master system of the robot performs the actual control, taking the planned motion into account while at the same time monitoring the real-time image information provided by OCT. It is important to note that during this stage, the surgeon is situated at a distance from the patient without any physical connection, thus maximizing the stability of the robotic system to make micrometer precision injections possible. Once the indicated needle tip target position

is reached, correct positioning can be confirmed with iOCT. At this point, the liquid can be injected and the progress monitored through the inraoperative visualization.

4. Finalization After successful delivery of the fluid subretinally and successful verification through imaging, the surgical procedure is finalized by retracting the robotic injector and removing it from the surgical field. The trocars are removed and normal post-intervention procedures are applied.

3.3 Path Planning and Execution

The robotic path planning and intervention execution component is, for the purposes of the visual guidance discussed in this thesis, considered an external component. It uses the semantic annotations provided from the periprocedural planning stage to compute a safe movement path of the robot. This involves avoiding the risk structures at a safe margin while approaching and reaching the indicated target position without destructive forces to the retinal tissue. The execution of the planned path in *shared control* mode provides suitable feedback, potentially using motion scaling and restricting the motion of the instrument tip to stay within the planned trajectory. In the *supervised autonomy* scenario, the robotic controller follows the planned path at a speed controlled by the surgeon, which allows them to intervene at any point in case critical deviations from the planned trajectory are detected.

4 Visual Planning and Guidance for Robotic Surgery

Contents

In this chapter, we present design concepts that were developed to evaluate and explore the visualization components required for an the integrated system like the one described in the previous chapter. These vertical design exploration prototypes served to gather experience with the various technical and visual aspects of the complex visualization systems.

Overview

To support the workflow of perioperative planning, the user interface offers an interactive review and annotation step during the procedure. We discriminate between the planning user interface and the procedure interface as these two stages have different user interaction requirements and visualize different data. During the planning phase, the main goal of the visual guidance system is to adequately show the static high-resolution OCT volume acquired for planning. The surgeon needs to identify the critical regions and needs to understand the robot access pathway and its limitations. In this stage, we need to provide a user interface to allow the surgeon to specify the target area for the injection as well as the risk areas, i.e. structures that must be avoided during the injection. During the procedure itself, the intraoperative guidance component needs to show the planned trajectory and the robot's actual position in relation to this path for the surgeon to verify everything is going according to the plan. From the visualization, the surgeon should be able to grasp the robot's current position and next movement direction in relation to the annotated risk areas and the target. Deviations from the plan should be immediately visible to allow timely intervention by the surgeon if necessary. Figure 4.1 depicts an overview of the high-level components and how they interact with each other in the system. The complexity of the visual guidance component is highlighted by the fact that it needs to interface to process the information from most other components and in order to visualize them in real-time.

4.1 Visual Planning

The visual planning component is integrated into the visualization user interface and provides a dedicated environment to plan the robotic motion by collecting the available information and fusing it into a single view.

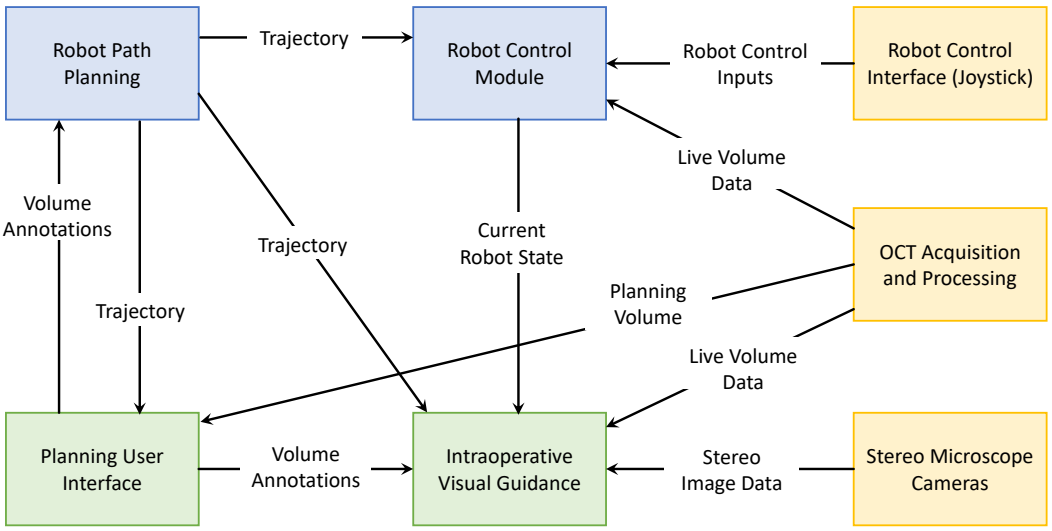


Fig. 4.1. Communication between the high-level components of the visualization prototype system.

It uses the volumetric image acquired before entering this phase and combines it with a frame from the camera image. It communicates with the path planning module to provide an interactive interface to define and modify the planned trajectory in accordance with the anatomical structures.

This section discusses work partially performed in the following thesis project:

T. Bařtipán. "A Visual Planning System for Robotic Subretinal Injection". *Advisors: J. Weiss, N. Navab. Master's Thesis, Technical University of Munich, 2019, Munich, DE*

Visualization Components

The screen is divided into a 3D viewport and a 2D view with additional UI elements. A representative screenshot is shown in Figure 4.1. The 3D viewport features a volume rendering of the planning volume data in combination with the label map that is defined as part of the planning. The two labels of the label map for risk areas and target area are shown in red and green overlays, respectively. The volume rendering is combined with an retinal enface image obtained from the color cameras of the microscope. The volume and camera image are co-registered such that the enface image in the background provides additional context on the retinal morphology. The volume rendering can be sliced with an axis-aligned clipping plane to show deeper layers of the tissue. We have found that it is helpful to review the 3D structure of the labels without the surrounding OCT volume and thus included a clipping mode that only cuts away the OCT data and not the label volume. The MPR view on the right is linked to the position of the clipping plane, which shows the respective slice in grayscale with the label maps overlayed. In addition to the label map showing target and risk areas, the visualization includes the trajectory that is required for the robot to reach the planned target position from

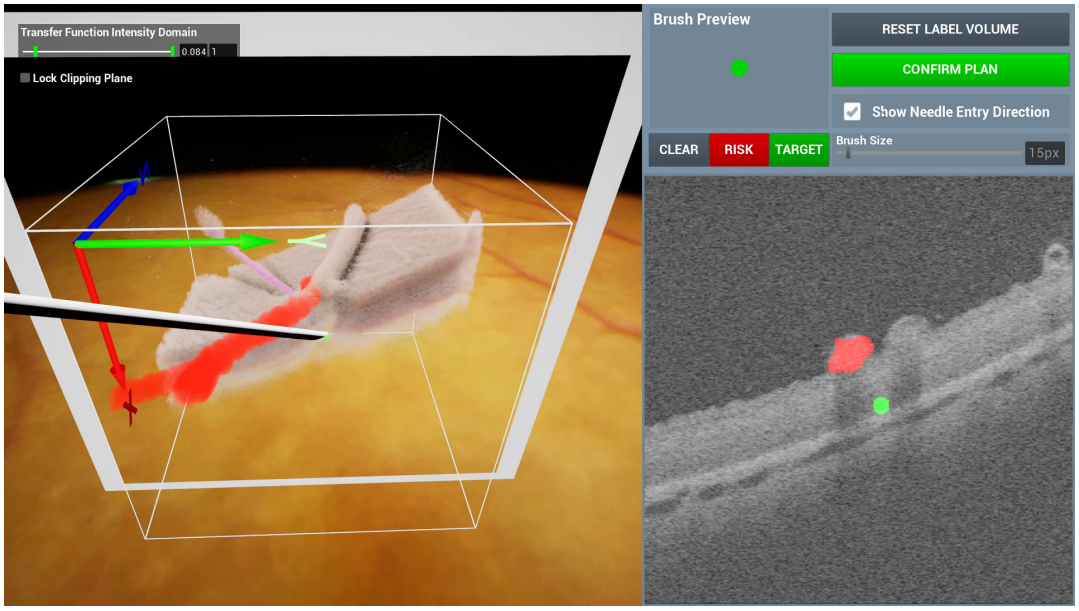


Fig. 4.2. Screenshot of the visual planning prototype, consisting of an interactive 3D view of the OCT volume as well as risk and target labels (left) and a 2D MPR view. The MPR view on the right is linked to the position of the clipping plane (gray frame, left). Parts of the blood vessels have been marked as risk areas (red); green dot and 3D model of the needle show the planned target point and trajectory.

its current position. Alternatively, the 3D rendering can be switched into a mode where it will display the grayscale data at the clipping plane directly.

Interaction

The 3D viewpoint can me manipulated using traditional pan-rotate-zoom gestures mapped to right mouse click and drag. The clipping plane follows the camera rotation so it will always clip along the principle axis that is facing the camera. This way users do not have to manually adapt the positioning of the clipping plane beyond the position along its principle axis, which is done via the mouse wheel.

Labeling of the main important information, the risk areas the robot must avoid and the target area the robot should reach, is performed via painting with a brush in either the MPR view or directly in the 3D view (which will map the mouse position to the corresponding intersecting point on the clipping plane). Controls for brush size and type are provided as well as the option to (partially) remove current labels. Modifying the risk or target areas will communicate those changes to the robotic path planning module and show the updated trajectory that results from this change, making an interactive workflow possible. Confirming the plan will switch the system into the intraoperative guidance mode, for which a concept will be described in the next section.

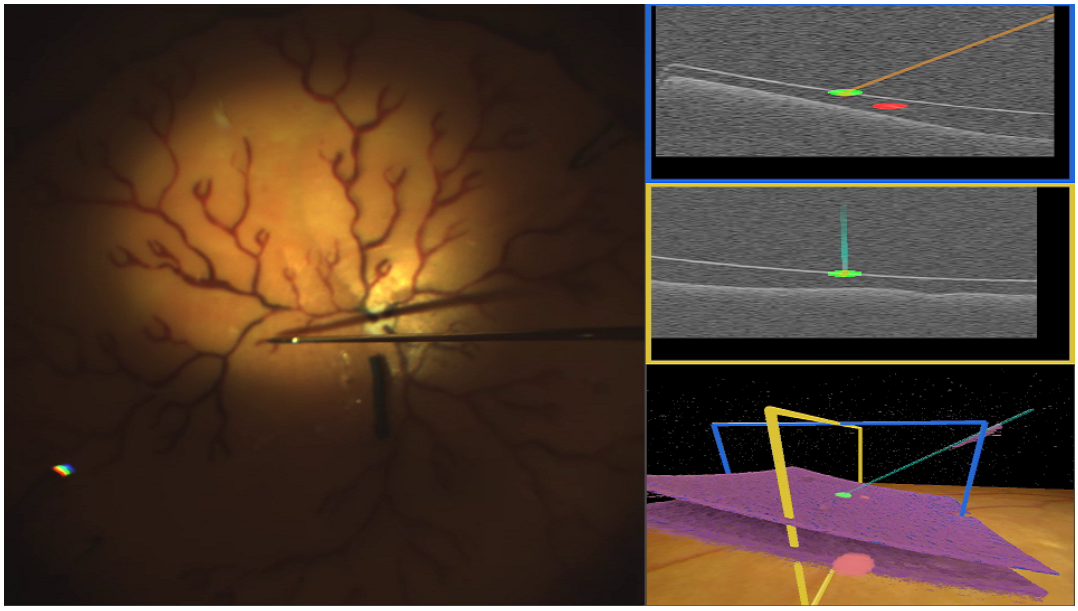


Fig. 4.3. Screenshot of the stereoscopic visual guidance prototype. Needle is on a straight trajectory towards the target point (green). From the 3D view in the bottom right it is apparent that the trajectory (turqouise line) matches the current instrument position (purple line visible in the top-right of the 3D view). Images obtained from an experiment in a plastic eye phantom with higher OCT reflectivity, which causes the DVR rendering to lose some of the surface detail.

4.2 Visual Intraoperative Guidance

During the execution of the robotic motion, we show the surgeon a combined view of live microscope image and live OCT imagery, combined with the information obtained from path planning and robot controller.

This section discusses work partially performed in the following thesis project:

N. Mirchev. "Intraoperative Visual Guidance for Robotic Eye Surgery". *Advisors: J. Weiss, N. Navab. Bachelor's Thesis, Technical University of Munich, 2019, Munich, DE*

Similar to the planning interface, the view is separated into a main viewport and a side view. The main viewport provides the live view of the stereoscopic camera feed from the surgical microscope and a right pane which shows the live OCT and planning information. Both parts of the screen are rendered stereoscopically, which enhances spatial understanding of the structures displayed.

The navigation view on the right side two MPR planes aligned relative to the current instrument pose: the upper plane is coaxial with, the second plane perpendicular to the current instrument direction. Depending on the stage of surgery, they can either be centered on the moving instrument tip or locked to the the target location to provide spatial context around the

instrument. The two MPR views also show the risk and target areas from the planning step as colored overlays over the OCT data. In addition, the planned trajectory is visualized in these slices. Even though the MPR views show mostly 2D information, they are actually rendered as stereoscopic 3D viewports which allows us to convey the trajectory as an actual 3D path relative to the plane. When viewed stereoscopically, the paths therefore give the impression of coming out of the image plane. To further enhancne the spatial understanding, the trajectory is rendered with a color map that encodes the distance from the MPR plane, going from blue (closer to the viewer) to orange (in-plane) to purple (behind the plane).

The 3D viewport on the bottom right of the view shows the live 4D OCT image combined with the trajectory as well as indicators for the locations of the two MPR planes. The live 4D OCT volume stream (at very low temporal resolution in this prototype due to the limitations of the FD OCT engine) is presented as a volume rendering. Using the known scan location of the planning OCT relative to the current live acquisition location, we can correctly relate the planned trajectory as well as the label volume for risk and target areas.

5 Robotic Surgery in Augmented Virtuality

Contents

Through the introduction of high-quality real-time digital microscopy with stereoscopic cameras and screens, a move towards digital systems has already been initiated in the surgical community. The addition of teleoperated robotic actuators in microsurgery marks an interesting change in the dynamics of the surgical room: with this setup, there is no technical reason anymore for the surgeon to be physically close to the patient as direct monitoring can be performed by a nurse or assistant. After the main surgical setup is performed and the robot is set up, it is possible to perform the intricate parts of the intervention that make use of the robotic system from a more remote position like a console or separate control room that receives all sensor input and controls the robotic end effector from there. This is similar to how some other teleoperated surgical robots, for example the da Vinci System (Intuitive Surgical), already operate. Such a fully digital environment presents an ideal opportunity to introduce mixed reality into the system as a way to further improve the visualization and control capabilities of the system. In addition it is easily possible to replace the conventional input keyboard, mouse and screen with a head-mounted display (HMD) and handheld spatial input devices.

A customized immersive environment, tailored to the intervention at hand, can offer several advantages over a traditional system:

- the available "real estate" for visualization in a virtual environment is much larger than on a traditional (stereoscopic) display and the virtual environment can be partitioned and adapted much more flexibly
- stereoscopy and motion parallax (user's head motion) reinforce depth perception and thus make it easier for the user to understand the complex geometry of the iOCT visualizations
- handheld motion controllers are in many situations more intuitive interaction devices in 3D environments than the mouse/keyboard combination, leading to simpler interaction paradigms and more precise control for the user

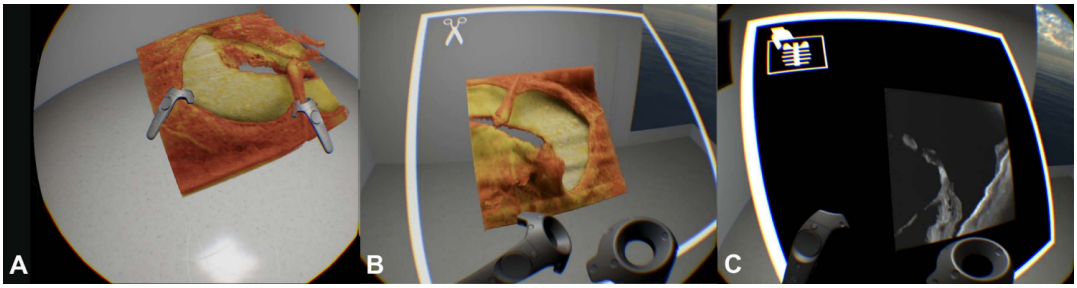


Image source: [138] CC BY NC ND

Fig. 5.1. VR environment presented by Maloca et al. [138], showing different visualization/interaction modes. **(A)** Bimanual grab and zoom gestures. **(B)** Positioning of arbitrary clipping planes. **(C)** Placement of arbitrary cross-sectional planes to inspect different aspects of the data.

In the following sections, we will present a design concept for such an immersive system and describe a concrete realization of this concept to support the perioperative planning stage of the intervention.

5.1 Related Work

One of the earliest works in this space is the system described by Hunter et al. in 1993 [94]. They describe a remarkably comprehensive teleoperated system for anterior eye surgery with bidirectional communication pathways for stereoscopic visual, audio and mechanical information. Their system relays stereoscopic visual information from a motorized pair of cameras that follows the operator's head motion. They combine this information with rendered virtual environment that is based on preoperative scans of the face and anatomy as well as intraoperative sensing of mechanical deformations of the anatomy. In addition, their system performs assistive functions such as motion scaling, tremor filtering and force feedback.

Most of the related research since then focused more on surgical training in virtual environments instead of directly supporting teleoperated systems. This body of work summarized well in the review by Khalifa et al. [105]. Later systems for surgical training include the surgical simulator prototype by Dumortier et al. [47] which featured a virtual environment observed through oculars similar to an actual microscope, manipulated via two haptic feedback devices. A commercialized version, the *Eye Surgery Simulator* (HelpMeSee Inc.) [88], features realistic rendering and a physics simluation to produce accurate haptic feedback and behavior for cataract surgery training. Based on a similar hardware platform, Cotin et al. [37], developed a retinal simulator with accurate FEM simulation of the retinal layers.

In 2015, Roodaki et al. [173] proposed the first augmented reality system that leverages the real-time information from OCT imaging to provide augmented feedback on tool-tissue distances. Their system detects instrument cross-sections in real-time in intraoperative OCT B-scans and visualizes the instrument's distance to the retina through a colored overlay on the OCT image. The augmented images are then displayed in the eye-piece of the microscope to provide direct feedback for the surgeon.

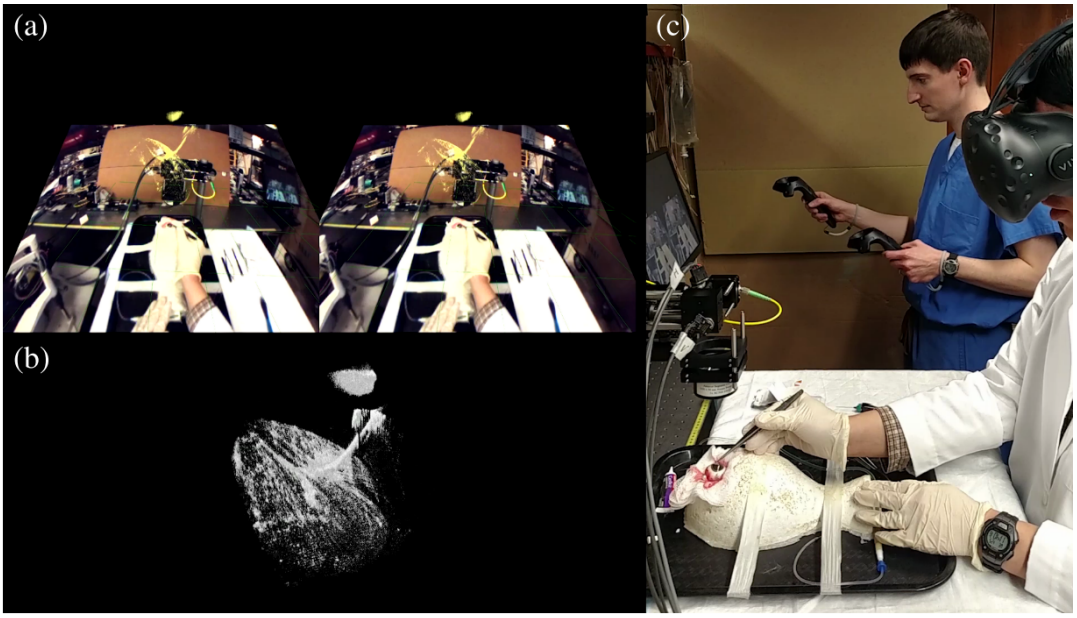


Image source: [43]

Fig. 5.2. Live OCT Viewer by Draelos et al. [43]. **(a)** Stereoscopic VR view, showing volume rendering of the live OCT data (yellow overlay) over the video passthrough image of the headset. **(b)** Isolated 4D volume rendering. **(c)** External view of mock surgery showing operator wearing VR headset and assitant manipulating the 3D volume-rendered view view.

In 2017, Dutra-Medeiros et al. [48] reported the first six successful vitreoretinal surgeries performed using a head-mounted 3D display that would show the stereoscopic camera feed of the microscope. In a comparative review of clinical experince with 3D display technology (including heads-up displays as well as head-mounted displays) in 2019 [148], Moura-Coelho et al. find both technologies to be promising alternatives to the standard microscope with their major benefits being improved ergnonomic setup for the surgeon as well as opportunities for teaching and communication as assistant surgeons and support staff can appreciate the same view as the operating surgeon.

Maloca et al [138] presented a high-fidelity immersive VR system for viewing and inspecting volumetric OCT data using an HTC Vive VR system. Their system includes basic bimanual interaction with the 3D cube as well as arbitrary cutting planes and manipulation of the virtual light source. In a study with 57 healthcare professionals and ophthalmology domain experts, they found that the vast majority found the experience agreeable and would use the system again to prepare for surgery. Using the same commodity virtual reality system (an HTC Vive, HTC Corporation), Draelos et al. [43] present a VR OCT viewer that supports inspection of not only recorded 3D OCT data but also 4D OCT data acquired in real-time during surgery. They demonstrate the feasibility of the system during a mock eye surgery, augmenting the OCT rendering on top of the video passthrough of the headset. Notably, in their discussion they already bring up the notion of VR-supported surgery, stating that "*VR-OCT may eventually become standard for intrasurgical navigation and medical education in OCT-guided ophthalmic procedures*" [43]. We share the same vision, however we think that VR technology can only create meaningful experiences when combined with robotic manipulators.

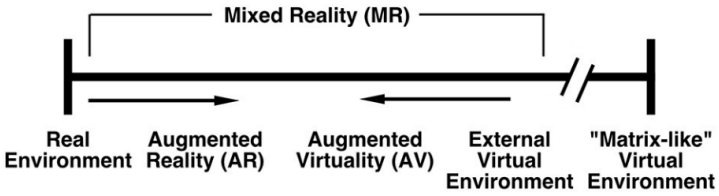


Image source: [189]

Fig. 5.3. Reality-Virtuality continuum according to Skarbez et al. [189], ranging from the real, unmediated reality on the left to a perfect virtual environment that is indistinguishable from reality on the far right.

5.2 Augmented Virtuality

The Reality-Virtuality (RV) continuum introduced by Milgram and Kishino in 1994 [145] provides a widely accepted framework to classify mixed reality experiences. The revised model presented by Skarbez et al. [189] distinguishes five principle levels of experiences (c.f. Figure 5.3), ranging from a real, unmediated environment to a perfect virtual environment that is perceptually indistinguishable from the real world for the observer. With respect to visual mixed realities, the most commonly encountered classes on this continuum are *Augmented Reality* (AR) applications (the real world is enriched with virtual elements) and *External Virtual Environments* (the external visual stimuli of the user are fully virtual), often simply termed *Virtual Reality* (VR). Augmented Virtuality (AV) is a third class inbetween these two where the environment is generally virtual but is augmented by elements from the real world.

Note: Realities and the *Matrix*

Skarbez et al. argue that the *Virtual Environment* of Milgram and Kishino only considers external stimuli that are perceived through the five *exteroceptive senses* (vision, hearing, touch, smell and taste). Even with a system providing those percepts as indistinguishable from reality, the *interoceptive senses* (notably the proprioceptive and vestibular senses) might still conflict. For this reason Skarbez et al. propose an additional class of perfect virtual environments at the extreme of the continuum, noting that these would have to be achieved by somehow overriding the interoceptive senses through direct brain stimulation.

In the context of providing clinically relevant information to a surgeon during a live surgery, a fully virtual environment that does not incorporate real information would be of limited usefulness. An AR approach on the other hand, enabled for example by an optical-seethrough HMD like the Microsoft Hololens, will allow the system to provide additional information in a relatively unobtrusive manner. However, this information will have to be shown *out of place*, for example by showing the current imaging data and robot motion plan floating within the OR. This can be an effective way to show additional information without being tied to a fixed screen position, however these augmentations will compete with real-world objects for space and contrast, making the experience less optimal. Furthermore, a big opportunity

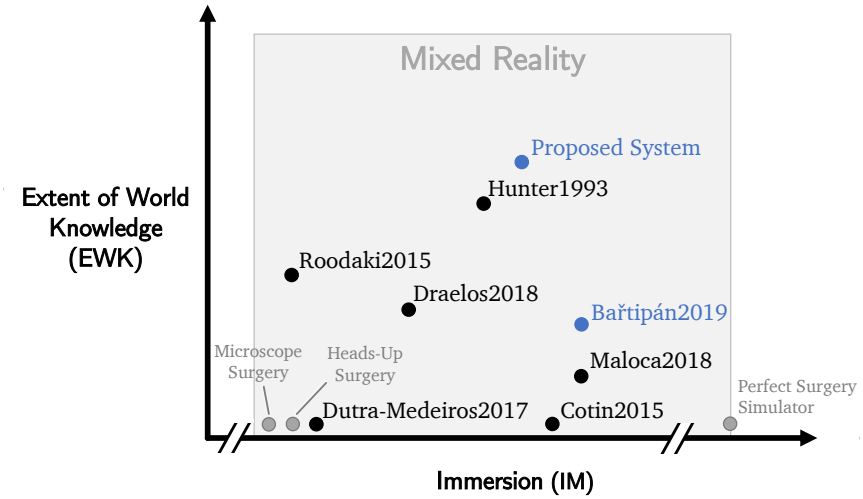


Fig. 5.4. Qualitative classification of related work and our work on Skarbez et al.'s 2D space spanned by Immersion (IM) and Extent of World Knowledge (EWK) axes [189]. Bařtipán2019 references the VR planning system presented in section 5.3, *Future System* the system discussed in section 5.4.

of AR applications, showing information *in-situ*, cannot even be leveraged in this situation due to the small scale of the real-world structures to be enhanced - the true trajectory of the needle could not be meaningfully augmented onto the patient's eye as it would only be a few millimeters in size. This would only be possible with an AR system that is integrated into the microscope image, which would again tie the viewer to a screen or the microscope oculus.

Instead, we argue that in this situation, an increased level of mediation will provide a better experience, and we thus propose an *Augmented Virtuality* (AV) system to support the surgeon. A fundamentally virtual environment provides a "clean canvas" to arrange the necessary information freely around the virtual observer without real-world limitations like screen sizes, mounting constraints or the presence of other physical objects like the microscope base or the operating table. Since the surgery is naturally indirectly observed and the robot also provides a means of indirect manipulation, all information that would be conventionally displayed on OR screens can also be shown in the virtual environment and can also be controlled from it.

In Figure 5.4, we qualitatively organize the existing works into the framework proposed by Skarbez et al. [189] for a better overview. The axis of *Immersion* (IM) corresponds to the RV-continuum discussed above, while *Extent of World Knowledge* (EWK) captures how well the system understands the real world around it. In this classification framework, the bottom left area represents essentially unmediated realities: Skarbez et al. use a pane of glass as an example, which is analogous to unaugmented surgery via the microscope. With increasing IM but low EWK we can classify systems that perform no active analysis of the surroundings, such as the video-passthrough display of Dutra-Medeiros et al. [48] or the surgical training simulator of Cotin et al. [37]. With higher EWK, the systems incorporate more information of the surgical scene into the experience: the system by Roodaki et al. [173], while not being particularly immersive, performs OCT image analysis to find instrument positioning and thus has more EWK than for example the system of Draelos et al. [43], which only provides an AR overlay of the un-enhanced OCT volume data.

In the following, we will introduce two novel systems that we have already classified in the above figure. The next section will describe *Bařtipán2019*, an immersive VR system to streamline the periprocedural planning step. In section 5.4, we will provide an outlook on how a future VR interface for robotic retinal surgery could look like.

5.3 Virtual Surgery Planning Environment Prototype

A major part of the periprocedural planning step consists of creating and reviewing the positions of the risk and target areas in the planning OCT volume. As a manual 3D labeling task, this is known to be tedious and time-consuming to perform accurately with traditional mouse/keyboard interactions, as the task essentially comes down to labeling each individual slice of the volume. The tracked motion controllers of commercial VR systems offer a native 3D interaction mode with which the task can be performed much more directly than with 2D input devices. Advances in automated semantic labeling and OCT segmentation using machine learning, for example the works the author has contributed to [192, 207, 239], can certainly simplify this part of the pipeline further. However, even reviewing and adjusting auto-generated labels will benefit from the improved spatial unterstanding and interaction offered by VR.

This section discusses work partially performed in the following thesis project:

T. Bařtipán. "A Visual Planning System for Robotic Subretinal Injection". *Advisors: J. Weiss, N. Navab. Master's Thesis, Technical University of Munich, 2019, Munich, DE*

5.3.1 Virtual Interaction Elements

The tracked motion controllers are represented as hands in the virtual environment. We use the standard VR interaction paradigm of grabbing objects with the handheld controllers in order to reposition and reorient them. This way users can flexibly rearrange the working space and inspect any object from all sides. Objects placed in the air will continue floating at this position irrespective of "correct" physical behavior. Beyond this, the software platform used to implement the prototype (Unreal Engine 4) allows for interactive floating 2D user interfaces that can be interacted with using a virtual laser pointer.

At the heart of the virtual environment is the interactive direct volume rendering of the OCT data used for planning. The data is rendered with a fixed transfer function, however the limits can be adjusted via a floating widget. Realistic shading is an important cue for structural perception of the data with DVR [41] and we have found this to be critical for OCT data. We use the algorithm by Sundén and Ropinski [199] to provide realistic lighting with

Fig. 5.5. Screenshot of the VR Planning application, showing intuitive volume annotation of risk areas using the tracked HTC Vive controllers, represented as hands in the virtual space.

multiple light sources using a precomputed illumination volume. We provide two directional light sources that can be interactively modified. Rotating either the OCT cube or the light sources will consistently update the illumination in real-time, thus providing a valuable tool for inspection of specific regions of interest. The volume rendering will also display the label volume in red and green with correct compositing with the OCT volume by adding the voxel intensities of both volumes together. The label rendering is not affected by the light sources and is rendered without shading and a constant opacity which leads to a glowing appearance of the labels. Figure 5.5 shows the resulting visual effect.

A clipping plane can be used to slice away arbitrary parts of the OCT volume, leaving visible only the labels of the risk and target area. The clipping plane is attached to the OCT cube, thus repositioning the OCT cube will also move the clipping plane such that the part that is clipped away from the volume stays the same. A floating window provides an MPR slice of the volume. The position of the MPR slice in the volume is linked to the position of the clipping plane, whereas the floating window itself can be repositioned arbitrarily and can be used as a secondary view for more details.

Labeling of the volume can be performed with either motion controller by pressing a button that will change the controller from the standard *manipulation mode* into a *labeling mode*. In this mode, a small sphere attached to the virtual hand will indicate the type of label to be drawn and the size of region, essentially defining a spherical volumetric brush. The brush size can be increased or decreased with two controller buttons and the brush type (risk area, target area, eraser) can be cycled through via another button. Assigning labels is performed via the trigger button, which allows for contiguous strokes while the button is pressed and held. While the user is labeling with one hand, they can still modify the clipping plane, light source position, or even the volume position at the same time with their other hand. This builds an intuitive and straightforward bimanual annotation experience.

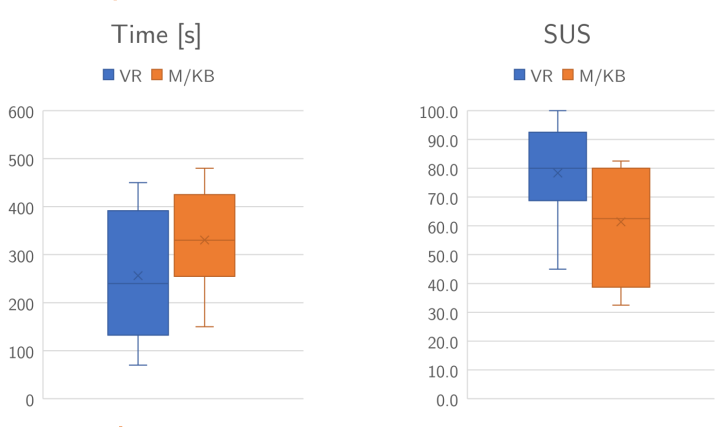


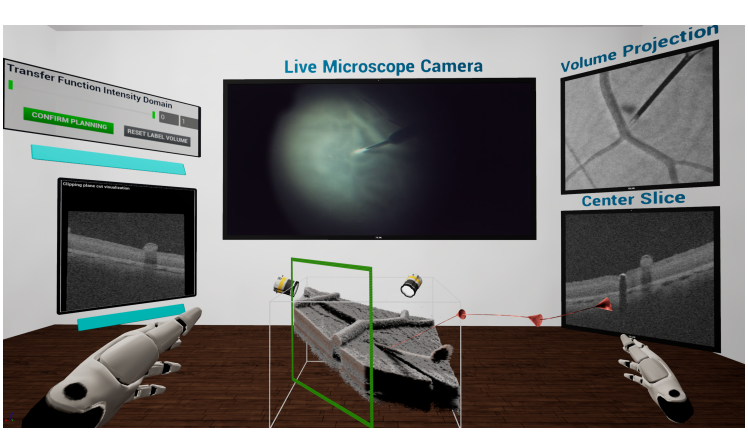
Fig. 5.6. Results of our small-scale study (n=9) comparing the VR system with the mouse/keyboard (M/KB) system detailed in section 4.1. Task completion time in seconds (left) and System Usability Score (right).

5.3.2 Preliminary Evaluation

A small-scale user study was performed to compare the VR-assisted planning process with the equivalent mouse/keyboard (M/KB) system described in section 4.1. Participants were tasked to label an identical OCT volume with both systems, by marking the superficial vessels as risk areas (c.f. Figure 5.5). Prior to starting the task, each participant was given up to 5 minutes to familiarize themselves with the task and the respective user interaction. The order in which the systems were used was randomized between participants. We measured task completion time as the time until participants felt satisfied with their labeling. After each system, we asked the participants to fill out a System Usability Score (SUS) questionnaire [21] to assess general feedback about the usability of each system.

A total of 9 participants (4f/5m, age 20-29 years) were recruited from students and staff of the Department of Informatics of the Technical University of Munich. No user exceeded the alotted 5 minutes time provided for familiarization with either system. On average, users took 5:30 min ($SD = 106s$) to complete the task with the M/KB setup and 4:15 min ($SD = 136s$) with the VR system, which corresponds to a 22.4% relative improvement of VR over M/KB. SUS scores were on average 61.4 ($SD = 19.96$) for M/KB and 78.3 ($SD = 16.58$) for the VR system. The metrics are plotted in Figure 5.6. Paired t-tests on both metrics revealed that the difference in time between the two systems was not significant ($p = 0.1412$) whereas the difference in SUS scores is statistically significant ($p = 0.0329$).

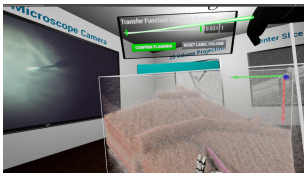
While this study cannot speak for the direct utility of the system in an OR setting over a conventional 2D system, it provides evidence that such a system can have interesting benefits in usability and potentially lead to lower task completion times at the same time. Oral comments made by the participants also suggest that, while the VR system is generally straightforward and less cumbersome, the M/KB version could have benefits when high precision is to be achieved as drawing on the MPR planes allows for more precise annotation of the structures.



(a) Proposed VR Cockpit Layout

Image source: ©AIRBUS S.A.S

(b) Cockpit of an Airbus A330



(c) First Person View

Fig. 5.7. VR Surgical Cockpit prototype. **(a)** Layout of the virtual room, showing stereoscopic microscope camera feed in the center, secondary OCT views on the right side and 4D OCT rendering on the left. **(b)** Similarities to an airplane cockpit with a main view surrounded by a plethora of information displays and controls. **(c)** Interaction with the volume data set using hand gestures.

From the perspective of the surgical workflow, the VR planning system described above is only of limited use in isolation: the effort of maintaining a separate room for the VR setup as well as the overhead of putting on the un-sterile VR equipment would be detrimental to the streamlined workflow of current retinal surgery if it were only used for this small part of the surgery. However, this planning application was developed as part of a future larger system for robotic VR surgery, which we will further describe in the next section.

5.4 Virtual Surgical Cockpit

The VR planning application has already shown that immersive VR has promising benefits when it comes to clear visualization and interaction. With the arguments for VR and AV environments discussed in section 5 in mind, we are convinced that extending this concept towards a fully immersive environment for robot-assisted retinal surgeries is an important next step towards effective and safe robotic interventions.

As a concrete example of how such a system could look like, we have implemented a concept which we call the *Virtual Surgical Cockpit*, drawing from the analogy of an airplane cockpit. Like in an airplane cockpit, a primary view (in our case the stereoscopic microscope image) is supplemented by additional information, all of which is needed to manage a complex machine that can be either steered directly or by an autopilot, which has to be supervised. Figure 5.7 illustrates the parallels between the two systems, showing a conceptual extension of the VR planning application to the VR cockpit.

The main view in our virtual cockpit consists of a stereoscopic image passthrough of the microscope camera images. This is analogous to how the 3D heads-up microscopes use physical screens to show the stereoscopic camera feed, however the virtual screen can be of arbitrary size. In addition, virtual screens can be used to present auxiliary information like an

average projection of the iOCT cube or MPR slices. It would also be easily possible to integrate patient vital signs or information from other devices in the OR if a digital interface is available to these devices.

We can also easily integrate the 3D rendering of the intraoperative data, combined with the data from the planning stage like the risk and target labels and the robot trajectory, as suggested in red in figure 5.7a. The DVR view of the volume data supports the same interactions as before, so the user can still optimize the rendering by adapting transfer function, cutting plane and light sources. This is especially important because the live imaging data has much more dynamic behavior and thus adjustments during the intervention are often required.

To control the robot, it is possible to provide the familiar joystick to allow for for direct control, replacing one of the VR motion controllers. Based on our own experience as well as feedback from the surgeons experienced with the robotic system, the joystick could be operated without the need of looking at it (which is prevented by the VR headset). However a representation of it could be incorporated into the virtual experience with the use of additional trackers if needed. An alternative control scheme could map the VR motion controller input to control the robotic motion directly. This could be an effective mode of control for the robot in semi-automatic mode, where the surgeon only needs to control the rate at which the robot shall follow the pre-planned trajectory to the target position.

Discussion

An interactive concept demonstrator has been realized based on the VR planning application. It can incorporate two camera streams from external USB cameras as well as recorded videos for demonstration purposes. It can also receive and display OCT imaging data via a simple TCP interface to interactively update the volume rendering. Interactive views related to the OCT imaging information (like the MPR slices and the enface projection) are computed in real-time via fragment and compute shaders directly from the OCT data.

While this setup serves to convey the concept of the *Virtual Surgical Cockpit*, a full realization poses significant additional technical challenges that would need to be fully solved for an acceptable experience: Firstly, the high bandwidth of OCT data is not trivial to transfer to the GPU efficiently without interfering with the tight rendering budgets imposed by a VR application. Especially with state of the art SS-OCT engines, texture transfers will need to be carefully managed in the background. Rendering the VR view at a stable, high framerate is an important factor in managing simulator sickness, therefore achieving stable frameraates will be a significant factor for user acceptance of such a system. The second challenge is interfacing with the robot control software. Careful calibration between all systems is required to avoid errors caused by mismatched coordinate systems. One solution for such a calibration was presented by Zhou et al. [239], leveraging the known shape of the surgical instrument. Lastly, introducing a VR system into or close to the operating room brings along logistical challenges. The technical setup for robotic interventions, while still in its infancy, is already quite involved (c.f. Figure 3.2). Finding sufficient room for a safe space for a VR setup, ideally in a standing configuration, will require careful consideration of the available space and workflow in the OR.

The opportunities for MR-assisted robotic microsurgery are vast, and we only described one of the many possible concepts for such an environment. Considering again the MR classification scheme by Skarbez et al. [189] (c.f. Figure 5.3), other systems could easily be derived: with decreasing IM, one could work with optical- or video-seethrough HMDs for more contextual awareness of the OR. Video-seethrough technology sounds especially interesting as it could adapt the video passthrough of the real world contextually based on gaze direction relative to the patient or considering the current workflow step. This can be a good middle ground to increase acceptance of the MR systems and pave the way towards more immersive environments: surgeons are often hesitant to directly enter a fully immersive experience during a live surgery. Another form of video-seethrough with less immersion would be augmentations on top of the microscope image, as is already done in a basic form with the commercial systems augmenting the OCT scan location and B-mode image into the ocular view, albeit with very low EWK. One axis that we have so far not considered in Skarbez et al.'s model is the aspect of Coherence (CO), i.e. how coherent the experience unifies sensory experiences. Highly coherent experiences react realistically to the actions and provide approriate sensory feedback. Extending the grab-and-move interaction to direct control of the robot, one could for example allow the user to actually grab the needle tip visible in OCT and push it towards the target position. Another concept for increased coherence could be to combine FEM simulation of the retinal tissue [37] with haptic feedback to allow the user to actually touch and deform the tissue virtually, possible gaining more insights into the structure and composition of the tissue.

6 Discussion

Developing the concepts presented in the previous sections as well as other quick prototypes as part of the thesis project was instrumental in understanding the problem space at the intersection of real-time visualization, OCT and robotic interventions. By limiting the scope of these projects to early technical prototypes, we were able to quickly learn about the limits of what is possible with the hard- and software platforms available today. This has gained us valuable insight into what a future platform will have to achieve, which can be distilled into the following required properties:

I. **High-performance OCT Imaging**: The OCT imaging platform that was available to us for these projects did not have the necessary imaging speed for full real-time volumetric imaging in high resolution. Without high A-scan rates, real-time 3D visualization suffers from poor resolution in either spatial or temporal domain, neither of which are acceptable intraoperatively.

II. **Efficient Processing**: The high data rates of OCT limit the amount of processing that can be performed, ruling out many image processing algorithms to improve or analyze the image data. Suitably efficient algorithms are required to sustain the real-time requirement of intraoperative visualization.

III. **Specialized Visualization**: Naive volume rendering of 4D OCT data is only of limited usefulness. Better solutions are required to visualize volumetric OCT data, as visualization techniques that work well with other imaging modalities are often not directly applicable to OCT imaging data because of different noise characteristics, imaging rates or the anatomical structures involved.

IV. **Semantic Awareness**: Contextual visualization needs to be improved by semantic information about the scene. In its simplest form, this can be semantic labels or layer delineations, to be taken into account by the rendering pipeline. Awareness of the robotic end effector and how it manifests in the imaging data are equally important to incorporate.

V. **Robotic Automation**: To fully realize the potential of VR-integrated robotic retinal surgery, the robotic platform needs to support advanced features like path planning, self-stabilization relative to tissue or automated path following.

VI. **End-to-End Integration**: Due to the high data rates and latency-sensitive nature of the robotic control signals, tightly integrated systems are required to minimize the communications overhead. Designing the communication infrastructure that binds together the different components is in itself a complex topic of research.

The research presented in the following chapters has been largely motivated by the vision of a workflow focused strongly on 3D visualization and, eventually, VR-integrated robotic surgery. Nonetheless, we often propose solutions to general problems in iOCT processing or visualization that are equally applicable to a more traditional setup like a 3D rendering on a secondary screen or a heads-up surgery system.

Part III

Methodology

7 Real-time 5DOF Instrument Tracking in OCT B-Mode Images

Contents

7.1 Introduction

To build better ditigal assistive systems for intraoperative use, especially considering AR applications, one important step is to integrate some awareness of the surgical scene. This information can come from explicit interfaces that provide status of connected systems like the mechanical state of microscope or OCT aquisition state. For more detailed information it is usually necessary to rely on and analyze the live imaging data in real-time. The high resolution, depth-resolved information of intraoperative OCT is an excellent source for accurate information, however due to its cross-sectional nature conventional computer vision algorithms that have been developed for natural (camera) images are not straightforward to adapt.

In this chapter, we present an efficient tracking algorithm that can determine the pose of a cylindrical instrument (such as an injection needle) with 5 degrees of freedom (DoF) from the cross-sectional information extracted from OCT B-scans in real-time. As an example of how this tracking information can be used, we show example application that provides the projected injection location as an AR overlay over the camera video image.

Sections 7.2 through 7.4 have been published previously in [216]. The last section 7.5 is based on the work presented in [218].

Key Contributions

- We show how the shadows and distinct reflection caused by the surgical instrument in the B-Scans are related to the incident angle of the tool and how this information can be integrated from several - not necessarily parallel - scans to infer the axis of the instrument.
- We derive an application specific Kalman filter that models the instrument movement between two acquired B-scans in order to tackle the latency between OCT scan lines.
- Our method can be applied to arbitrarily positioned B-mode patterns, supporting for example parallel, crossing or volumetric patterns.
- The utility of the algorithm is shown by an AR guidance application to show the projected injection point on the retina to assist subretinal injection.

7.2 Related Work

Prior work focused mainly on the use of microscopic RGB images. In this context, Richa et al. [169] presented tool tracking based on weighted mutual information between stereo images. If the 3D CAD model of the tool is known, the instrument pose can be recovered by projective contour modelling [8]. Sznitman et al. [202] classified each pixel as either background or tool part using a multiclass ensemble classifier. The precise localisation of different tool parts is subsequently obtained by a weighted averaging on the response scores. Allan et al. [2] estimate the full 3D pose based on a level-set algorithm incorporating optical flow. Rieke et al. [170] propose to combine a fast color-based tracker with a robust HoG feature-based 2D pose estimator via a dual Random Forest. An offline learning with online adaption approach further increased the generalisation regarding unseen backgrounds and instruments [171]. Supervised Deep Learning based approaches [68, 117, 118] require an extensive annotated dataset to capture the wide range of image distortions such as blur, specular reflections and limited focused field of view.

Despite the recent advances in instrument tracking based on microscopic RGB video, none of the methods can tackle a major inherent disadvantage: Even if the tracking precision is perfect in the microscope image, it cannot yield precise depth information through its projective imaging geometry. Acquiring information from the high resolution OCT at the 3D location of the instrument would still not be feasible if the accurate spatial mapping between the microscopic image and the iOCT is unknown. Although optical microscopy and iOCT can share the same optical path in a device, this alignment requires complex calibration routines. For the same reason, traditional navigation solutions such as optical tracking or electromagnetic tracking are not applicable as they usually have an accuracy in the range of 200 to $1400\mu m$. The intraoperative OCT on the other hand has an axial resolution of 5 to $10\mu m$, which is close to histopathology. Therefore, we propose to track the 5DOF pose of the surgical instrument directly in the iOCT B-Scans and by that completely avoid the bottleneck of calibration.

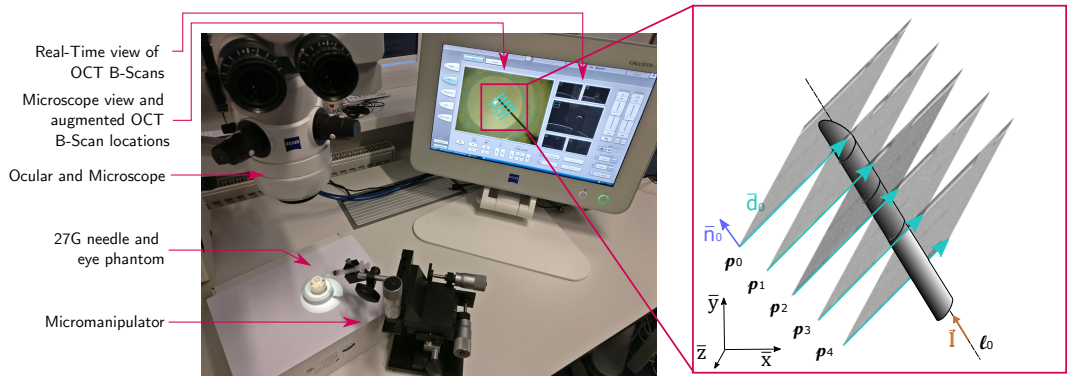


Fig. 7.1. **Test scenario and coordinate system.** *Left:* We use an OPMI Lumera 700 surgical microscope together with a RESCAN 700 iOCT system and a Callisto Eye assistance system, all from Carl Zeiss Meditec, Oberkochen. *Right:* Explanatory sketch describing notation and spatial relationships between B-Scan direction and needle.

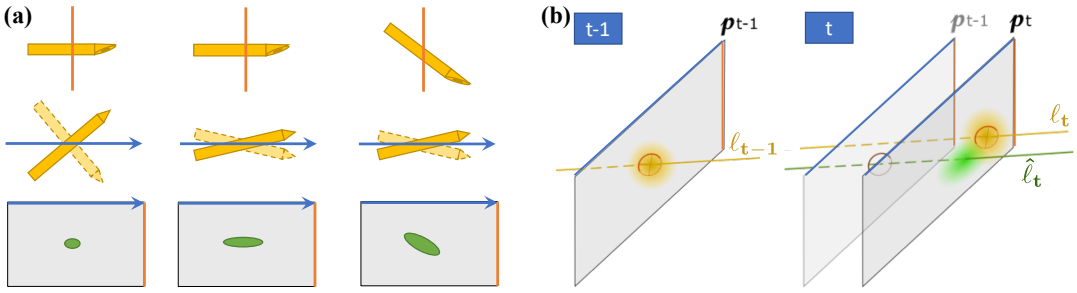


Fig. 7.2. **Geometric Modelling. (a)**The cross-section of the tool on a single B-Scan allows to determine its 3D axis ℓ up to one ambiguity. **(b)** The Kalman filter resolves this ambiguity and relates between time steps: A linear motion model of time step $t-1$ gives a line prediction $\hat{\ell}_t$; this is corrected by the ellipse parameters measured at time t and leads to final estimate ℓ_t. The blurred positions in image planes ρ indicate the estimated error covariances.

OCT is a fundamentally different imaging modality than conventional microscopic imaging: It measures echo delay and intensity of back reflected near-infrared light waves [50]. Instead of an RGB en-face view of the surgical scene, the B-Scans provide grayscale, cross-sectional information. Due to the underlying physics, conventional surgical instruments appear hyperreflective and consequently any signal beneath them is lost (c.f. Figure 7.3(b)). A first step towards instrument tracking in this type of data was proposed by Zhou et al. [238] in terms of instrument segmentation based on a fully convolution network. The method however requires a volumetric dataset, is not applicable to real-time applications and is restricted to static instruments.

7.3 Tracking Instruments over Time

This section details the methodology for our tracking algorithm, explaining the geometric setup, ellipse detection and fusion of the data in the newly derived extended Kalman filter.

Intuition: Temporal Tracking

The key idea that our algorithm is based on is that the intersection shape of the instrument in a single B-mode image already contains much of the relative pose of the instrument. Under the assumption that the instrument does not move (or equivalently, subsequent B-scans are acquired *instantly*), we could directly determine the instrument axis by connecting the ellipse center points between the ellipse centers of two B-mode scans. Because the time difference between B-scans is significant with respect to the expected motion, we use the Extended Kalman Filter as a way to integrate both the time delay between B-scans as well as instrument motion into a consistent model.

7.3.1 Geometric Setup

OCT B-mode imaging is generally performed by scanning the sampling beam over the tissue in a repeated pattern. The A-scan samples from linear segments along this path are combined into a B-mode image (B-scan). In the case of ophthalmic surgery, motion of tissue and instruments are slow enough that the samples of a B-scans image can be assumed to be at the same point in time. Additional dead-times between B-scans however mean that the motions during typical surgical maneuvers can result in visible motion between subsequent B-scans. For interventional OCT imaging, usually a fixed pattern of several parallel and/or orthogonal scanlines is used to provide the surgeon with different cross-sectional views (B-Scans) of the working volume. From the known layout of this pattern, a transformation to 3D space can be computed for each pixel. Commercially available iOCT devices typically have a fixed A-Scan rate of 27-35 kHz, resulting in a B-Scan update rate of 27-35 Hz if 1000 A-Scans per B-Scan are assumed. The number and placement of B-Scans is determined by scanning patterns which can be flexibly interchanged during an intervention. To set a reference coordinate system, we define the axis of positive horizontal deflection as our x axis and the vertical deflection as the y axis. As the projective field of view for small B-Scan lengths is narrow, we can neglect the slight projectivity of the system and assume a euclidean coordinate system instead, thus assuming that the z axis is parallel to our A-scan direction. Figure 7.1 illustrates the geometric relationships.

Note: Notation

We use bold lowercase symbols, e.g. $\mathbf{x}$, to symbolize column vectors and bold uppercase symbols, e.g. $\boldsymbol{K}$ for matrices. Vectors that represent a position/direction in 3D space are denoted with $\vec{x}$ and normalized 3D vectors as $\bar{x}$. The first derivative with respect to time is denoted by $\dot{x}$.

The plane corresponding to a B-scan in 3D that is reconstructed from each scanline is parametrized as $\rho : (\vec{x} - \vec{p}) \cdot \bar{n} = 0$ where $\vec{p}$ corresponds to the top-left corner of the

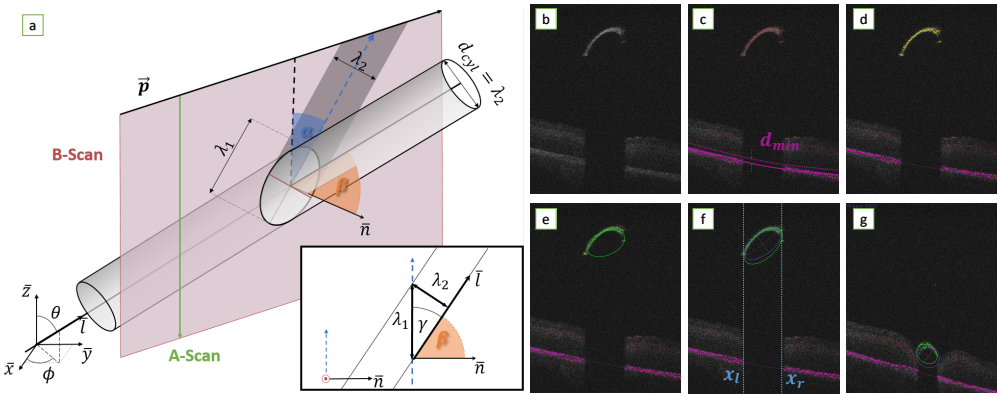


Fig. 7.3. **Ellipse parameters and detection.** **(a)** Schematic view showing the relationship between ellipse parameters and the needle. **(b-d)** Steps during ellipse detection: **(b)** Input image. **(c)** Candidate points (violet + red) and fitted tissue layer (violet) with inlier margin. **(d)** Tool candidate points p^*_{tool} (yellow) obtained by filtering and thresholding $d(p)$. **(e)** RANSAC-fitted tool ellipse and inliers (green). **(f)** Final ellipse obtained by non-linear optimization (purple) and parameters. **(g)** Example that our method is still able to detect the ellipse even if it is touching the tissue.

B-Scan image and $\bar{n}$ is the plane normal. As a simplification, we model the tracked needle as an idealized cylinder with known diameter d_n and an axis parametrized as

$$\ell : \vec{x}(\tau) = \vec{x} + \tau \begin{pmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{pmatrix} = \vec{x} + \tau\bar{l}, \tau \in \mathbb{R},$$

where θ and ϕ are azimuth and polar angle of the axis direction.

Note: Spiral Scanning OCT

We did not consider non-linear scanning such as spiral acquisition modes [25] in this work given their more recent introduction. To integrate our algorithm with these modes, one could still interleave spiral scanning with linear patterns for the purposes of instrument tracking, or extract resampled B-mode planes from the spiral.

7.3.2 Ellipse Detection for Orientation Estimation

To calculate the axis of the needle in 3D, we use the cross-section of the tool that is visible in an OCT B-Scan $B = (b_{r,c}) \in \mathbb{R}^{n_r \times n_c}$ with n_r rows and n_c columns. It can be shown that the intersection between a cylinder and a plane in the non-degenerate case forms an ellipse, which we parametrize through the center position in image coordinates $(C_x, C_y) \in \mathbb{R}^2$, the length of the long (λ_1) and short (λ_2) axis and the angle α that is formed between the longer axis of the ellipse and the negative y-axis of the image (or equivalently, the angle between λ_1 and $\bar{z}$, c.f. Figure 7.3**(a)**). Only the hyper-reflective surface of the metal needle is visible

on an OCT frame while everything below is shadowed. Our algorithm for ellipse detection and determination of its parameters is performed through the following steps, illustrated in Figure 7.3**(b-f)**.

1. **Candidate Points** are defined as the pixels with maximum intensity along each A-Scan (column) of the B-Scan image: $p^c_{cand} = (\mathrm{argmax}_r(b_{r,c}), c)$. The set of all candidate points over all colums is p_{cand}. These generally correspond to either the tool's surface (p_{tool}), noise (p_{noise}) or an anatomical layer (p_{eye}) such as the corneal surface in anterior segment or retinal pigment epithelium in posterior segment images.

2. A **Tissue Layer** representing the global eye shape is fitted based on these candidate points by using the RANSAC algorithm. For anterior surgery and posterior surgery with non-degenerate RPE, a circular model is used. In cases where a circular model is not suitable, we use a polynomial of order 4 to approximate the anatomical layer more closely.

3. **Tool Candidate Points** $p_{tool}*$ can now be separated from the anatomical surface p_{eye} by computing the distance of every candidate point to the tissue model $d(p)$. To remove isolated candidate points that belong to p_{noise} or not-hyperreflective tissue layers, we apply a 1D morphological opening, closing, followed by median filtering each with a 15 px kernel and obtain a filtered distance list $d^*(p)$. The tool candidates are then defined as the set $p_{tool}* = \{p \in p^c_{cand} | d^*(p) > d_{min}\}$ where d_{min} is a threshold to remove points close to the surface.

4. *(optional)* For **suspected pathologies** (for example due to preoperative imaging), we iteratively exclude from $p_{tool}*$ all points which are closer than d_{min} to an already excluded point, effectively removing all points which are *connected* to the tissue layer.

5. A **First Ellipse Estimate** is computed by fitting an ellipse to the set of tool candidate points $p_{tool}*$ using RANSAC, provides an inlier set p_{tool}.

6. **Ellipse Refinement** is achieved by minimizing the geometric distance between the ellipse and the points p_{tool}. We directly assign $C_x = 0.5 * (x_r + x_l)$ where x_l, x_r are the x coordinate of the leftmost and rightmost point of p_{tool}, and suppose a known needle diameter d_{cyl} to set $\lambda_2 = d_{cyl}$.

7. **Tool center line** ℓ_t can then be related to the ellipse parameters by:

$$\begin{array}{ccccc} \cos\alpha & = & \bar{z}\left(I - \bar{n}\bar{n}^T\right)\bar{l} & = & \cos\theta, \\ \frac{\lambda_2}{\lambda_1} & = & \cos\beta & = & \bar{n}\cdot\bar{l} \end{array} \tag{7.7}$$

   where β is the angle between the cylinder axis and the plane normal (c.f. Figure 7.3**(a)**). The first relationship follows by considering the projection of $\bar{l}$ into the B-Scan plane, computed as $\left(I - \bar{n}\bar{n}^T\right)\bar{l}$. Since α is the angle between the A-Scan direction $\bar{z}$ and this projected vector, its cosine is equal to their dot product, which directly simplifies to $\cos\theta$. The second equality is developed from considering the inset view of Figure 7.3(a): From

the right triangle shown, $\sin\gamma = \frac{\lambda_2}{\lambda_1}$ follows, and $\sin\gamma = \sin\left(\frac{\pi}{2} - \beta\right) = \cos\beta = \bar{n} \cdot \bar{l}$ from trigonometry and the dot product definition.

7.3.3 Extended Kalman Filter

From the ellipse parameters in one single cross section, it is not possible to uniquely reconstruct the pose of the cylindrical needle. We use a Kalman filter [98] to fuse the noisy measurements from frames at different time points and infer the current pose of the needle in each frame. This section develops the state and measurement transition of the extended Kalman Filter (EKF) that nonlinearly filters our measurements.

Intuition: Kalman Filters

Kalman filtering is a form of statistical filtering that can estimate a *hidden variable*, e.g. state that is only observed indirectly via (noisy) measurements. We can use such a filter if we have a model for how the hidden variable behaves over time (the state transition) an second model for how the measurements result from the state (the measurement transform or observation model). With these known forward models, a new measurement can then be propagated backwards to update the hidden state variable. In our case, the state is the pose of the instrument and the measurements are the parameters of the ellipse. We assume an acceleration-based dynamic model for the instrument, so we also include velocities in the state.

The Kalman filter requires modeling of a *state transition* and the *measurement transform* based on the previous state $\mathbf{x}_{t-1}$, previous control vector $\mathbf{u}_{t-1}$, process noise $\mathbf{w}_{t-1}$ and measurement noise $\mathbf{v}_t$:

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}, \mathbf{w}_{t-1}) \qquad \textit{state transition} \tag{7.8}$$

$$\mathbf{z}_t = h(\mathbf{x}_t, \mathbf{v}_t) \qquad \textit{measurement transform} \tag{7.9}$$

State Representation

The state and knowledge about our system are represented by the state vector $\mathbf{x}$ and control vector $\mathbf{u}$,

$$\mathbf{x} = \left(\vec{x}^T, \theta, \phi, \dot{x}^T, \dot{\theta}, \dot{\phi}\right)^T,$$
$$\mathbf{u} = \left(\bar{n}^T, \vec{p}^T, \Delta t\right)^T,$$

where $\vec{x}, \theta$ and ϕ model the needle axis and $\dot{x}, \dot{\theta}$ and $\dot{\phi}$ model its current velocity and angular velocities, respectively. The control vector contains the parameters defining the iOCT plane of the next measurement as well as the time since the last measurement.

State Transition

We assume that our tool is influenced by unknown accelerations and angular accelerations $\mathbf{a} = (\vec{a}_{\vec{x}}^T, a_\theta, a_\phi)^T$ drawn from a zero-mean Gaussian distribution. Our preliminary state update is thereby defined as

$$\begin{pmatrix} \vec{x}_t^* \\ \theta_t \\ \phi_t \\ \dot{x}_t^* \\ \dot{\theta}_t \\ \dot{\phi}_t \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 & \Delta t & 0 \\ 0 & 0 & 1 & 0 & 0 & \Delta t \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \cdot \mathbf{x}_{t-1} + \mathbf{w}_t, \tag{7.10}$$

where $\mathbf{w}_t = (\frac{\Delta t^2}{2}\vec{a}_{\vec{x}}^T, \frac{\Delta t^2}{2}a_\theta, \frac{\Delta t^2}{2}a_\phi, \Delta t\vec{a}_{\vec{x}}, \Delta t a_\theta, \Delta t a_\phi)^T \sim \mathcal{N}(0, \mathbf{Q})$.

However, as the ellipse measurement is related to a different image plane at the next timestep, we move the base point of the line x_t to that plane (c.f. Figure 7.2 (b)). Therefore, we determine the final update for $\vec{x}_t$ by intersecting the predicted tool center line $\hat{\ell}_t$ defined by $\vec{x}_t^*, \theta_t$ and ϕ_t with the image plane ρ_{t-1} of the next measurement which is defined by the control vector $\mathbf{u}_{t-1}$. Solving the intersection between $\hat{\ell}_t$ and ρ_t results in a nonlinear state transition for $\vec{x}$:

$$\vec{x}_t = \hat{\ell}_t \cap \rho_{t-1} = \vec{x}_t^* + \frac{(\vec{p}_{t-1} - \vec{x}_t^*) \cdot \bar{n}_{t-1}}{\bar{l}_t \cdot \bar{n}_{t-1}} \cdot \bar{l}_t \tag{7.11}$$

With Equations 7.10 and 7.11, our state transition function is fully specified. Re-basing the line onto the next OCT plane allows the Kalman filter to retain low error covariance for the position of the line due to the resulting simple measurement transition, as opposed to letting the base point of the tool line be arbitrary and performing the line-plane intersection as part of the measurement transition. The more complex update of the position implies that the new position is non-linearly dependent on the process noise variable $\mathbf{a}_t$. Therefore, we use the formulation of the EKF with non-linear noise to update the predicted error covariance matrix as

$$\boldsymbol{P}_{t|t-1} = \boldsymbol{F}_{t-1}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F}_{t-1}^\top + \boldsymbol{L}_{t-1}\boldsymbol{Q}\boldsymbol{L}_{t-1}^T \tag{7.12}$$

where the $\boldsymbol{P}_{t-1|t-1}$ is the estimated error covariance matrix of the previous time step and the Jacobian matrices of the state transition function $\boldsymbol{F}_{t-1} = \frac{\partial f}{\partial \boldsymbol{x}}\Big|_{\hat{\boldsymbol{x}}_{t-1|t-1}, \boldsymbol{u}_{t-1}}$ and $\boldsymbol{L}_{t-1} = \frac{\partial f}{\partial \mathbf{w}}\Big|_{\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{u}_{t-1}}$ are derived analytically.

Measurement Transition

From a single B-scan we find the parameters of the cross-sectional ellipse as described above. We define the measurement transformation

$$\hat{\mathbf{z}}_t = h(\mathbf{x}_t) + \mathbf{v}_t = \begin{pmatrix} \vec{c}_t, \\ \cos\alpha \\ \frac{\lambda_2}{\lambda_1} \end{pmatrix} + \mathbf{v}_t \tag{7.13}$$

where c_t are the 3D coordinates of the measured ellipse center and the measurement noise $\mathbf{v}_t \sim \mathcal{N}(0, \boldsymbol{R})$ is assumed as additive Gaussian noise. Above equation is motivated by Equation 7.7, together with the fact that $\vec{c}_t$ is on the tool axis and we can thus set $\vec{c}_t = \vec{x}_t$. The other components of $\mathbf{v}_t$ are not directly related to $\hat{\mathbf{z}}_t$. The standard EKF innovation equation is $\mathbf{S}_t = \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^T + \mathbf{M}_t\mathbf{R}\mathbf{M}_t^T$. Again, we are able to determine the Jacobian $\mathbf{H}_t = \frac{\partial \hat{\mathbf{z}}_t}{\partial \mathbf{x}}\big|_{\hat{\mathbf{x}}_t}$ and $\mathbf{M}_t = \frac{\partial \mathbf{h}}{\partial \mathbf{v}}\big|_{\hat{\mathbf{x}}_{t|t-1}} = \mathbf{I}$ analytically.

Due to our choice of representation and parametrization, two mathematical singularities arise: Elements of F_{t-1} and H_t tend to infinity for $\theta_t \to k\pi, k \in \mathbb{Z}$ as well as for $\ell_t \cdot n_{t-1} \to 0$. The first implies that the needle is parallel to the A-Scan direction while the second represents the needle parallel to the OCT plane, so for both cases we are not able to see elliptical cross sections in the OCT frame. We avoid numerical instabilities in our predictions by only updating the predicted state if an ellipse has been detected and increasing the time step Δt for the next frame when no ellipse could be detected.

7.4 Experiments and Results

In this section, we evaluate the performance of our algorithm regarding different aspects: We first discuss computational performance and how we estimated the measurement noise covariance matrix. Then we analyse the algorithm in a series of experiments on both anterior and posterior segment in phantom eyes and ex-vivo porcine eyes in terms of movement stability and robustness to pathologies. Finally, we demonstrate an application example of our algorithm which consists of an injection guidance application.

Parameters and Initialization: For the minimum distance of candidate points to the tissue layer, we set d_{min} to 20px, which corresponds to 50microns. The measurement noise covariance matrix $\boldsymbol{R}$ is determined by analyzing the covariance of the ellipse parameters across several data sets where the needle is at an unknown but fixed angle with respect to the B-Scans. Based on the physical interpretation of the process noise as an unknown acceleration induced by the surgeon (c.f. section 7.3.3), we empirically choose values for $\vec{a}_{\vec{x}} = 3.0mm/s^2$, $a_\theta = a_\phi = 60deg/s^2$ and derive $\boldsymbol{Q}$ based on the definition of $\mathbf{w}_t$. These parameters are used across all experiments. In all our experiments, we initialize the Kalman filter with $\mathbf{x}_0 = (\vec{x}_0, \theta_0, \phi_0, 0, 0, 0)$ with parameters of the line fitted through the centers of the first two detected ellipses.

Computational Performance: The prediction and estimation steps for the EKF reduce to matrix operations on matrices not larger than 10x10 elements, which can be implemented very

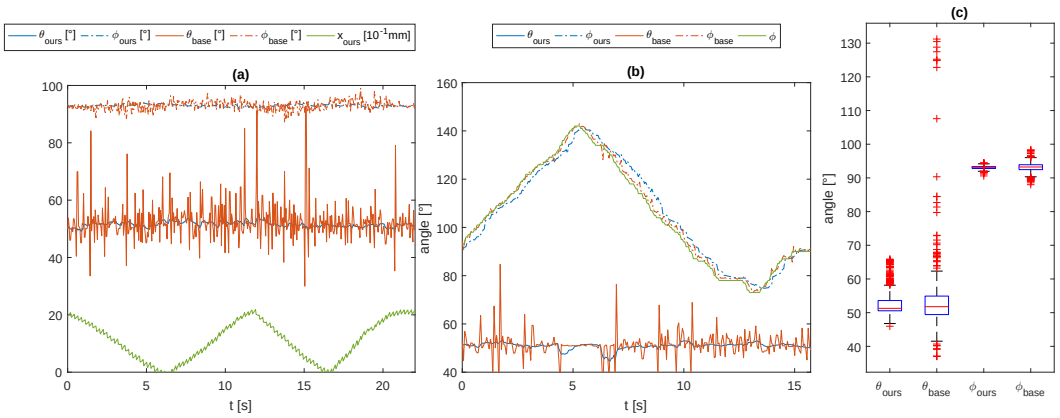


Fig. 7.4. **Phantom Evaluation.** Parameters with subscript $base$ are from line fitting while $ours$ indicate the proposed method. **(a)** Needle movement lateral to the B-Scan direction. Due to the fixed needle orientation, θ and ϕ are expected to be constant. The baseline method exhibits higher variation which, in the case of ϕ_{base}, is correlated with the lateral movement direction, while our method retains a more stable orientation. **(b)** Needle rotation around the z-axis simulated by rotation of the OCT scanning pattern (green). The known rotation angle ϕ is recovered robustly while our method shows better stability regarding the expected constant angle θ. **(c)** Box plot of estimated angles during axial movement. Our method shows much reduced variation and therefore better results regarding the reconstructed orientation.

efficiently. Therefore, the most computationally intensive part is the processing of each B-Scan to detect the ellipse and find its parameters. Our CPU-based native implementation (C++) with circle fitting and without pathology handling is able to process 1024x1024px B-Scans in under 5.4 ms (186 FPS), on a Notebook with an Intel Core i7-6820HQ CPU @ 2.70 GHz and 16GiB RAM. We are therefore are easily able to process the OCT framerates of current iOCT engines, which range around 27-32 FPS at this resolution.

Movement Stability Evaluation: We evaluate the movement stability of the proposed method on both phantom and ex-vivo porcine eyes. Since a comparison to optical tracking methods or other traditional methods is not feasible, we employ a mechanical micromanipulator for generating ground truth in terms of known, precise 3-DOF movements along an axis with fixed direction for the surgical tool. To evaluate the quality of our estimation, we compare our method to line fitting through the ellipse centers of two subsequent images.

Phantom Experiments: A first set of experiments was performed with a 27G needle in an otherwise empty field of view with an OCT scanning pattern of five parallel B-Scans to assess the stability of our algorithm to different kinds of movements. With the needle fixed at a constant orientation, we move it only along one axis of the micromanipulator in order to determine the influence of translations on the pose estimation. Figure 7.4**(a)** shows the effect of lateral motion on the estimated direction of the needle. It can be seen that our method greatly reduces the variations in θ. Furthermore, the baseline method confuses lateral movement as a change of angle, resulting in a visible correlation between ϕ_{base} and the lateral movement direction, which we indicate by red/blue bar (red means $v_x < 0$, i.e. movement to "left"). Our method does not exhibit this effect as it models not only position but also velocity. The same effect can be seen for needle movement along the Z-Axis (c.f. Figure 7.4**(c)**), where the movement influences the azimuthal angle θ of the baseline estimate. Variation

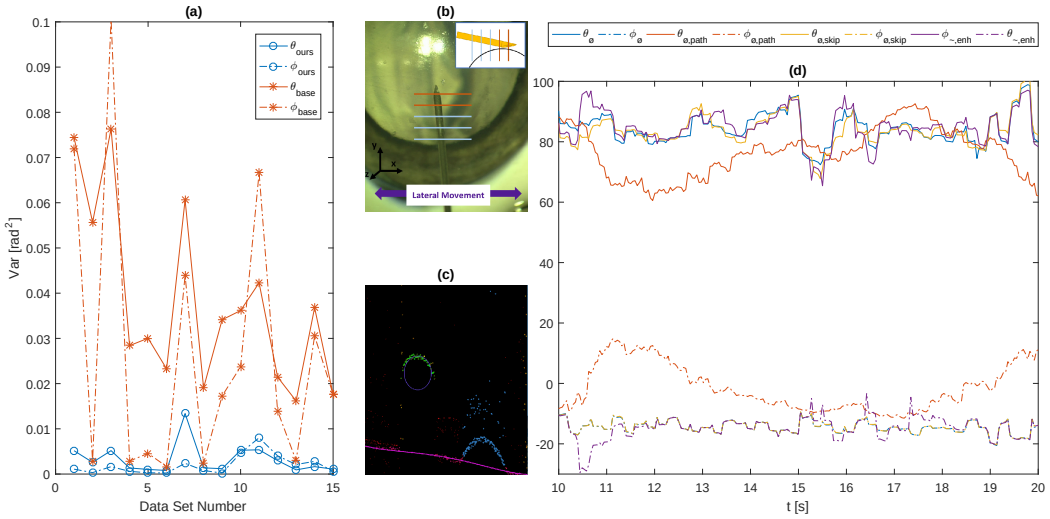


Fig. 7.5. **Ex-Vivo Evaluation.** Reconstruction of the needle orientation during lateral movement. *Movement stability:* **(a)** Analysis of the variance of the estimated orientation during lateral movement with fixed orientation. Our method shows reduced variance for both angles in all data sets. *Robustness to irregular tissue or ellipse detection failure*: **(b)** Robustness to failing ellipse detection is verified by simulating failed detection in B-Scans marked as red. **(c)** Polynomial fit and additional pathology detection (Step 4) can distinguish pathology candidates (blue) from ellipse points (green). **(d)** The circle fit ($\theta_\emptyset, \phi_\emptyset$) performs worse for the same movement if pathologies are present ($\theta_{\emptyset,\text{path}}, \phi_{\emptyset,\text{path}}$). A polynomial tissue model with pathology handling can reconstruct the needle orientation ($\theta_{\sim,\text{enh}}, \phi_{\sim,\text{enh}}$). Needle axis stability is also maintained when ellipse can only be detected in three of five B-scans due to the needle touching the tissue in the other scans ($\theta_{\emptyset,\text{skip}}, \phi_{\emptyset,\text{skip}}$).

in θ is higher compared to ϕ due to the higher uncertainty of ellipse detection in the axial than in the lateral direction. As we cannot produce precise rotation with the mechanical micromanipulator, we instead record an image sequence during which the OCT scan pattern is rotated around the z axis and then manipulate the metadata to ignore the known rotation, yielding an image sequence that is equivalent to rotating the tool around the z axis. The analysis of the fitted orientation in Figure 7.4 **(b)** shows that our algorithm is able to reliably reconstruct this rotation while being less susceptible to noise in the ellipse estimation of each frame. A slightly delayed angular adaptation of our method is noticeable due to the smoothing property of the Kalman filter. However, we argue that a rotation as strong as in this data set rarely occurs in ophthalmic surgery, where needle movement is generally very slow and controlled.

Ex-vivo experiment: To evaluate the transfer towards real scenarios, we performed a similar experiment on enucleated porcine eyes. We acquired a series of 15 anterior data sets from 5 different eyes, each with the same setup with a fixed needle angle and lateral movement. Figure 7.5**(a)** shows that our algorithm can still robustly estimate the translation and greatly reduce the variance in the estimated angle.

Irregular tissue and ellipse detection: To investigate the robustness of our method in more challenging cases, we have tested the following modifications on one of the ex-vivo data sets with lateral needle movement (c.f. Figure 7.5**(b-d)**): To simulate a needle being too close to the tissue to be found by our ellipse detection, we force the ellipse detection to fail in two of five B-Scans. It can be seen that our method is able to retain stable tracking. To simulate

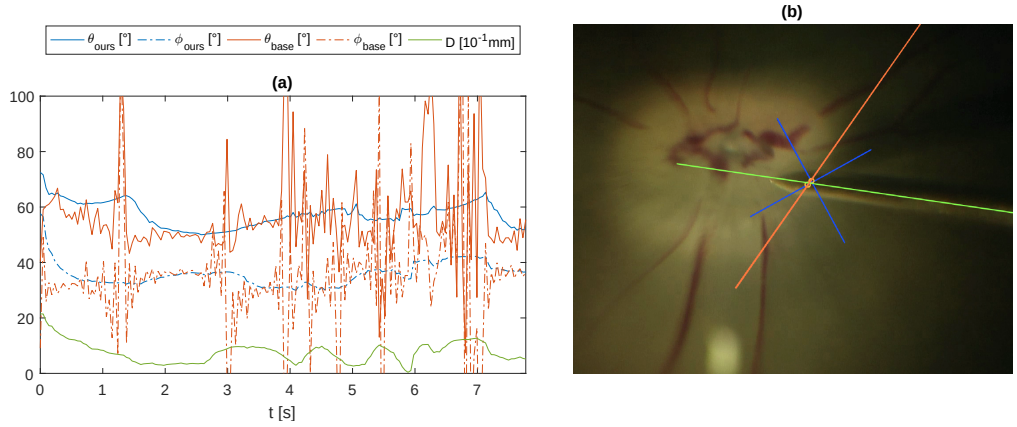


Fig. 7.6. **Freehand movement in ex-vivo experiment. (a)** Analysis of freehand needle movement in posterior segment while scanning a Cross-Pattern of two perpendicular B-Scans. D (green line) is the distance between estimated tool axis and intersecting line of the two B-Scans. Linear fitting fails to compute a reliable pose while our method can still provide a stable tracking. **(b)** Microscope view with OCT scanning location overlaid in blue. Yellow circles indicate the centers of the detected ellipse from the two B-Scans. Orange line is the estimated line from the baseline method. Green line is our estimated, highlighting the benefit of using the ellipse shape for more stable tracking.

pathologies, we shift the candidate points p_{cand} to resemble an irregularly shaped retina (Figure 7.5**(c)**). The experiment shows that our method using a circular tissue model alone $(\theta_{\emptyset,\text{path}}, \phi_{\emptyset,\text{path}})$ is problematic in the presence of pathologies. Thus, if they are expected from preoperative data (or the tradeoff can be generally accepted), enhanced pathology handling $(\phi_{\sim,\text{enh}}, \phi_{\sim,\text{enh}})$ can successfully recover stable tracking, at the cost of higher per-frame processing time to 7.1ms.

Pattern comparison and Freehand movement: We performed an experiment with freehand movement of the needle inside the OCT region while scanning with a pattern consisting of only two perpendicular B-Scans. Figure 7.6**(a)** shows the orientation of the sequence once again compared to the baseline method. This shows the baseline method being unable to provide a meaningful estimate when subsequent points are too close together, which is the case when the needle moves closer to the intersection of the two B-Scans (Figure 7.6**(b)**). It can be seen that our method is susceptible to bad initialization by the linearly fitted line through the first frames, however it is able to converge to a stable tracking after a few seconds and retain this pose even when the needle is close to the center. This demonstrates that our estimator can still infer the needle orientation from the ellipse shape when the ellipse centers alone do not provide enough information.

7.5 Visual Injection Guidance

As an example application, we have designed an assistance application that provides injection guidance during subretinal injection by showing the surgeon the projected intersection point of the tracked needle with the target layer. During an actual injection, the OCT would be optimally placed through this injection point to give the surgeon a good impression of the current needle depth. Manual repositioning is however not feasible. We use the proposed

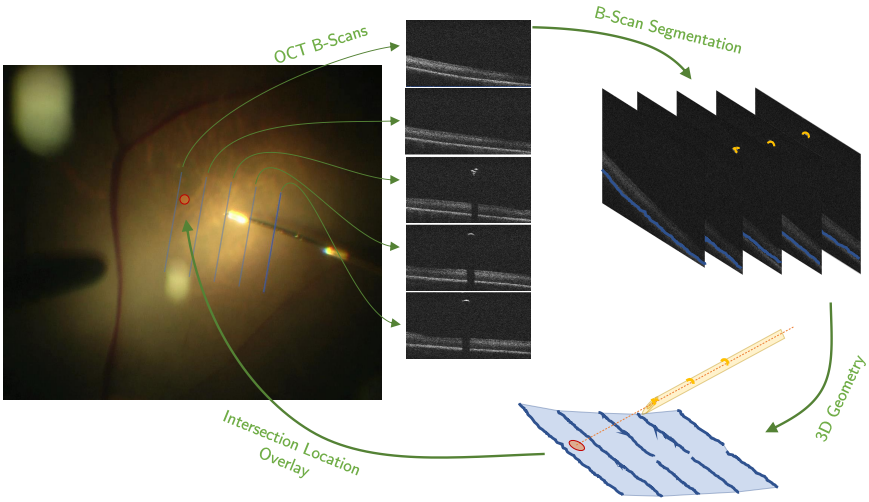


Fig. 7.7. **Screenshot of the injection guidance application.** Left: Augmented view of the surgical scene, showing the camera view with the overlaid OCT scanning locations as well as the projected intersection point with the RPE layer. Current and last B-Scan are marked with white and blue bars for illustrative purposes. Right: Schematic view of the 3D relationships between B-Scans (blue), current needle estimate (green), and intersection point with the target surface (red). These relationships cannot easily be inferred from a simple 2D microscope image.

algorithm to track the injection needle and show the projected injection point overlayed on the microscope image. To estimate the intersection point, we first reconstruct the target surface by using the tissue surface points p_{eye} of the ellipse detection stage of each B-Scan, which correspond to pixels on the RPE for posterior images (c.f. Figure 7.3**(b)**). We reproject the points p_{eye} from several B-Scans to 3D space and fit a sphere using RANSAC. The intersection points of the tracked tool axis ℓ with the estimated sphere are projected to the camera coordinate system using the 2D calibration provided by the manufacturer, which is valid in the current microscope focus plane. On top of the video image, we draw a circle corresponding to the needle thickness to indicate the injection point. Thus, we provide distance perception without explicitly tracking the needle tip, as the surgeon can infer the distance of the needle to the surface by the distance of the projected intersection point and the instrument tip visible in the camera image.

Evaluation: We evaluated our method on both anterior and posterior phantoms. In both scenarios, we acquired three sequences of free needle movement. For acquisition, we used a Zeiss Lumera 700 with RESIGHT 700 in 5-line HD OCT mode and the needle is constantly touching the target surface to have ground truth. Setting up the needle to touch the retinal tissue allows us to determine the injection position as ground truth from the video image. The OCT imaging was set up to be focused on parts of the needle that were sufficiently far away for our algorithm to detect the instrument axis. In the en-face camera view, we manually annotated the ground truth touching/intersection point in each image as the needle tip in 811 images and compare it against the position provided by the intersection computation. We report a median error of 0.230mm (mean: 0.299+-0.062mm) for anterior and median error of 0.268mm (mean: 0.358+-0.090mm) for posterior guidance. These errors are likely in part due to the inaccurate labeling process of the ground truth labels, however noise during tracking can cause some intermittent outliers. These are usually not critical as they only apply for one or two frames before the Kalman filter re-stabilizes.

7.6 Conclusion and Outlook

We presented a novel algorithm for tracking a surgical needle in 3D space solely using the high-resolution, cross-sectional OCT view. The method makes no assumptions on the layout of the B-scan scanning pattern and is therefore easily integratable into existing systems with dynamically changing scanning patterns. We avoid expensive computations by geometric modelling and consequently, our method is able to process more than 180 B-Modes / s on a 2017 notebook CPU. While this can be likely scaled up to support SS-OCT imaging rates with more recent desktop CPUs and a more efficient, multithreaded implementation, we believe that a major strength of our work is that it still works with only a few sparse B-Mode images, which could be interleaved with or extracted from the full scanning pattern. In experimental evaluation both on phantom and ex-vivo porcine eyes, we show that the method is able to compensate needle movement between subsequent B-Scans, thus exhibiting increased robustness against low B-Scan frequency and generating a more stable estimate compared to line fitting. We demonstrate the usefulness of our tracking algorithm by providing a simple augmented reality scenario for subretinal injection.

Note: Follow-up Work

At the time of publication (2017), this work was the only published OCT-based instrument tracking algorithm applicable to real-time iOCT. Since then, the author of this dissertation has contributed to two follow up works in this space: [239] and [192] both employ machine learning for more robust layer and/or instrument segmentation in the B-scans as a basis for the estimation of the instrument position. What still sets this algorithm apart is the explicit modeling of instrument motion between B-scans and its cheap computational cost without the need for GPU acceleration. These are important characteristics when only a limited computational budget is available for tracking, for example when integrating algorithm into a larger visualization system.

With regards to our robotic assistance system, this algorithm can be used as a basis for showing the current axis of the needle in a live 3D view. The AR overlay application could be extended to or integrated into a 3D view, showing the projected injection point relative to the planned target area. This information could even be fed back to the robotic control software to create a closed-loop targeting system that can compensate for movements of the eye or errors in the registration. Since we specifically designed the algorithm to be computationally efficient, it provides advances to the requirements of efficient processing (**II.**) and semantic awareness (**IV.**) we have defined in section 6.

Layer-Aware OCT Rendering

8

Contents

8.1 Introduction and Background

As we have discussed previously, OCT-based intraoperative visualization for vitreoretinal surgery is a highly relevant yet underexplored topic in scientific visualization. In the following, we will present two visualization concepts that are specifically designed to support decision making during surgery based on 3D OCT data. The methodology is designed for contemporary commercial intraoperative OCT (iOCT) systems which only have limited intraoperative 3D imaging capabilities: current commercially available iOCT systems have an A-Scan rate of 27-35 kHz, resulting in acquisition times of 1-2 seconds for a typical volume. While this is not suitable for real-time volumetric feedback, it enables a workflow where the surgeon pauses for the imaging to assess the surgical situation in a 3D view. In the context of robot-assisted surgery, this could be at the perioperative planning step, but it also applies in manual surgery where the surgeon might want to stop and review the current state of e.g. a peeling procedure. At the time of writing, these devices do not include any intraoperative visualization of this volumetric data beyond reviewing them slice-by-slice. Among other procedures, iOCT has been shown to be beneficial in ERM peeling procedures [54], a procedure that requires the removal of a thin membrane attached to the retinal surface. With the recent introduction of precise ophthalmic robotic surgery systems [172], visual feedback for robotic maneuvers during operations like subretinal injections are also of increasing importance.

An important anatomical reference is the Retinal Pigment Epithelium (RPE) layer, a thin, hyper-reflective layer below the retinal surface: in ERM peeling (c.f. section 2.1), distance from this layer is a reliable factor for structural analysis. In subretinal injections, the RPE is the most important boundary and should not be punctured while approaching it as closely as possible for optimal positioning of the injection bleb. We present a volumetric rendering concept that

Key Contributions

- We present a perceptually linear color map that encodes the positioning relative to retinal layers in chrominance and the OCT reflectivity in luminance.
- We develop a layer-aware DVR rendering using this color map that aid sturctural perception of the retinal surface.
- We introduce Layer-Adjusted Maximum Intensity Projection (LA-MIP), a novel projection that corrects for the natural curvature of the retina to aid subretinal injection tasks.
- We show how the DVR rendering can be adapted for integration with a 4D SS-OCT system in combination with a stereoscopic display to provide real-time volumetric feedback.

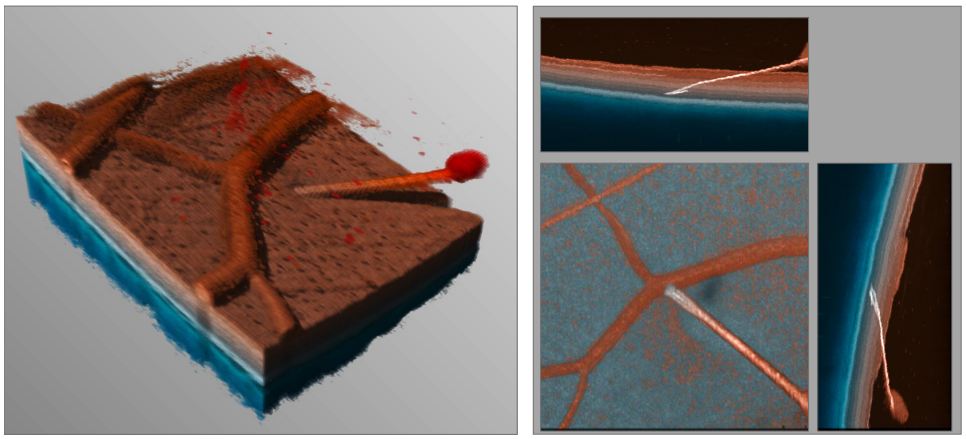
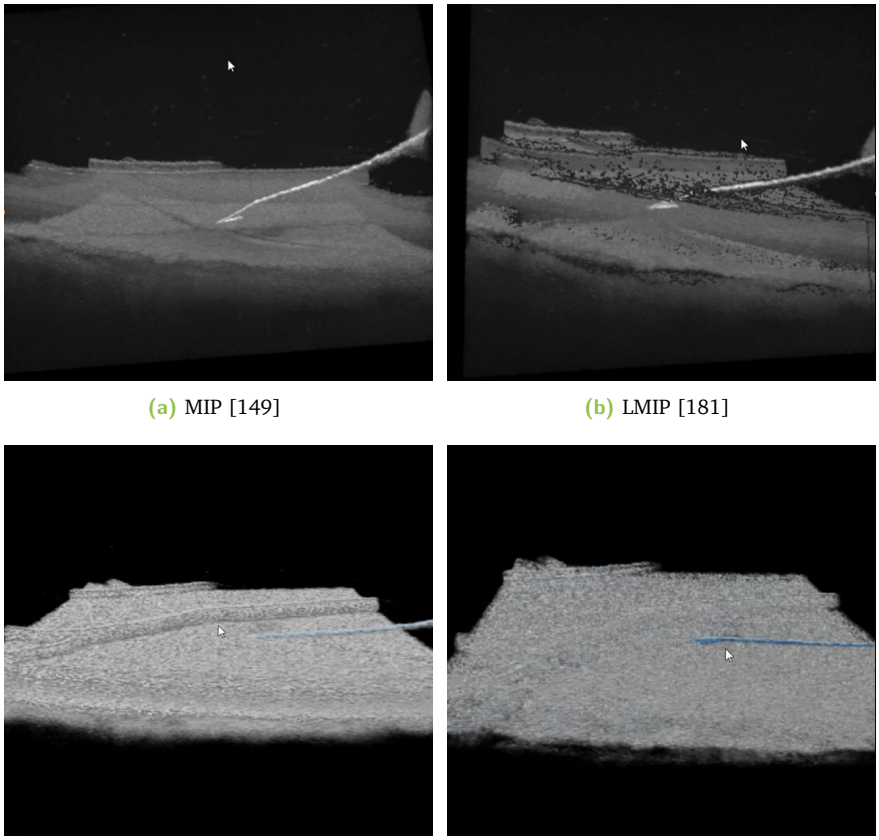


Fig. 8.1. Local shadowing together with a colorization relative to an anatomical reference layer provides clear visualization of epiretinal structures *(left)*. A colorized en-face projection combined with a novel *layer-adjusted MIP* provides improved visualization of instrument position below the surface *(right)*.

aids understanding of superficial retinal structures. It highlights small superficial structures through a combination of enhancement methods, making it well-suited for ERM peeling and other surgeries where perception of the surface details is important. We supplement this with a *Layer-Adjusted Maximum Intensity Projection* (LA-MIP) which extends axis-aligned maximum intensity projections of OCT data and allows for a better assessment of needle penetration depth during an injection.

Parts of this chapter have been published previously in [215]. Section 8.3.1 provides additional information about the prototyping phase preceding the main method that did not make it into the main paper. Going beyond contemporary SD OCT engines, section 8.6 shows how we have adapted the principle method to integrate with a high-performance SS-OCT engine after publication.



(a) MIP [149] (b) LMIP [181]

(c) DVR with Phong-Shading [44] (d) MIDA [22]

Fig. 8.2. Screenshots of conventional volume rendering methods applied to the same OCT volume, with comparable settings between all four methods. Images have been created using voreen (`voreen.uni-muenster.de`).

8.2 Related Work

With iOCT as a generic 3D volumetric data cube, all generic direct volume rendering approaches can be applied [22, 44, 149, 181]. However they are, for the most part, unable to provide sufficient visualization of internal structures, as illustrated in Figure 8.2.

In the current clinical practice, slice-based review of the data cube is the most common way surgeons interact with the volumetric data. In diagnostic systems OCT cubes are often reviewed via en face projections of *slabs*, where each pixel corresponds to the averaged intensities of an A-Scan between two tissue boundaries. Until recently, intraoperative volumetric OCT has been presented either on a secondary external screen, or through the use of monocular or stereoscopic screens integrated into the microscope [24]. In research setups, feedback of 3D visualization has shown promise and are seen as the way forward [57]. Viehland et al. [211] have developed a volume rendering method method specific to OCT volumes. They combine classic Phong-shaded Direct Volume Rendering (DVR) with enhancements for samples with high gradient magnitude, voxels with gradient perpendicular to the viewing direction as well as adding a slight colorization with increasing depth. They demonstrate improved perception of structures in OCT volumes. Bleicher et al. extend this approach with axial colorization of

intraoperative OCT volumes [18]. In their work, they apply colorization of the OCT based on the axial position relative to a re-computed center of mass. To our knowledge, the only rendering methods that are specifically proposed for OCT is by Viehland et al. [211] and its extension to axial colorization by Bleicher et al. [18]. This method has been impactful and is already used in several studies ([24, 111], among others). This led us to evaluate it in the context of our two use cases. In own experiments with different data sets, we have found their approach to be especially susceptible to noise.

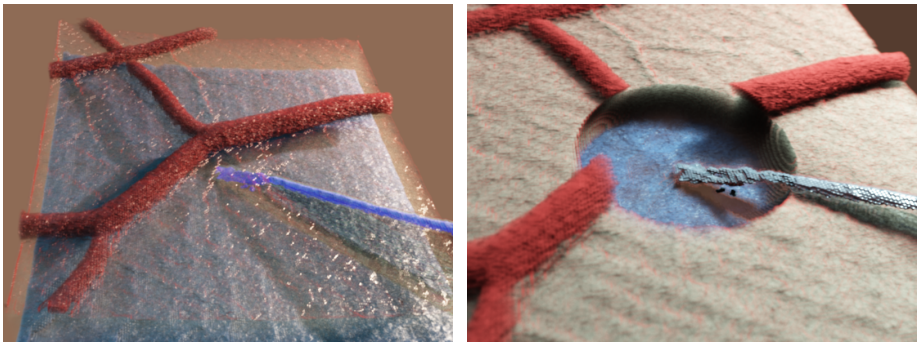
Viehland et al. mitigate this problem by applying a 3D Gaussian filter to remove much of the noise. However, this also has the effect of reducing visibility of the small structures we are interested in. In our own attempts to visualize data sets with their method with a lower amount of denoising, a good visualization could often only be found by turning off the gradient enhancement, and even then slight changes to the transfer function limits could have a high negative impact. The depth enhanced colorization method introduced in [18] has shown promising results for iOCT rendering by increasing the visual contrast and adding depth cues through colorization. However, their approach uses a fixed axial reference position for the colorization to enhance tissues at different depths. According to discussions with our clinical experts, differences in absolute depth position are not always of direct interest. Natural curvature of the retina leads to color differences which have to be differentiated from actual structural anomalies. Lastly, the method was specifically designed to provide enhanced visualization of superficial structures. However, in the case of subretinal injections, an accurate visualization of spatial relations below the retinal surface is needed.

8.3 3D Rendering of intraoperative OCT Data

8.3.1 Visual Prototyping with Monte Carlo Rendering

Developing volume good volume rendering visualizations for a specific modality is a complex and challenging task. A flexible, feature-rich platform is key to expedite the creative process of exploring possible visualizations interactively and adapt them from expert feedback. There are several frameworks built specifically for this purpose such as Voreen[1] or inviwo[2]. CAMPVis[3] [185], is a similar framework that has been used extensively for the development of the later methods in this chapter and section 9. However, to explore the level of required visual fidelity, we started with a volume renderer with a very realistic illumination model. As a basis for experimentation, we chose the publicly available implementation of ExposureRender[4], a GPU-accelerated implementation of physically-based Monte Carlo volume rendering implemented in Cuda. With a mid-range mobile GPU (NVidia GTX 960M for most of our experiments), the probabilistic rendering typically converges to a usable preview within 2-4 seconds, which enables an interactive workflow.

[1] http://voreen.uni-muenster.de
[2] https://inviwo.org/
[3] https://gitlab.lrz.de/CAMP/campvis-public
[4] https://github.com/ThomasKroes/exposure-render, [113]



(a) Semi-transparent surface with emissive needle (b) Contextual cutout centered on needle tip

Fig. 8.3. Visualization prototype with semantic labels enable a more fine-grained control over the visulization, if precise labels are available.

Note: Exposure Render

The freely available open-source version of ExposureRender offers the following tunable features, which were highly valuable for visualization prototyping:

- Freely adjustable camera position as well as physical camera model with aperture, exposure and focus parameters can simulate over/underexposure and (de)focus
- Appearance modeling through an intensity-based transfer function to control diffuse, specular and emissive components
- Arbitrary number of area light sources that can be adjusted in color, intensity, size and position
- Optional background illumination with adjustable color and intensity

In an effort to experiment with more semantically aware renderings, we adapted the implementation to support an additional per-voxel label volume, which allowed us to apply distinctly different transfer functions to structures. In addition, we implemented semantic cropping within a sphere or along specific cut planes, which would only remove specific labels. Figure 8.3 shows two such concepts: Figure 8.3a shows an emissive instrument with a thin, semi-transparent retinal surface a and red superficial blood vessels. Figure 8.3b shows a spherical cutout around the instrument tip, leaving the instrument itself as well as the RPE surface visible. With the semantic information available through the labels, reduced visualizations can be achieved that are almost illustrative in style. Even though these concepts generally elicited positive interest from our experts, we did not follow this direction further as by that point, intraoperative availability of semantic labels of the quality required seemed questionable.

Instead, we started developing concepts that solely rely the ILM and RPE layers as these are easier to segment. We further extended the implementation to support color maps based on position and , which allowed us to define color maps based on the distance to those two layers

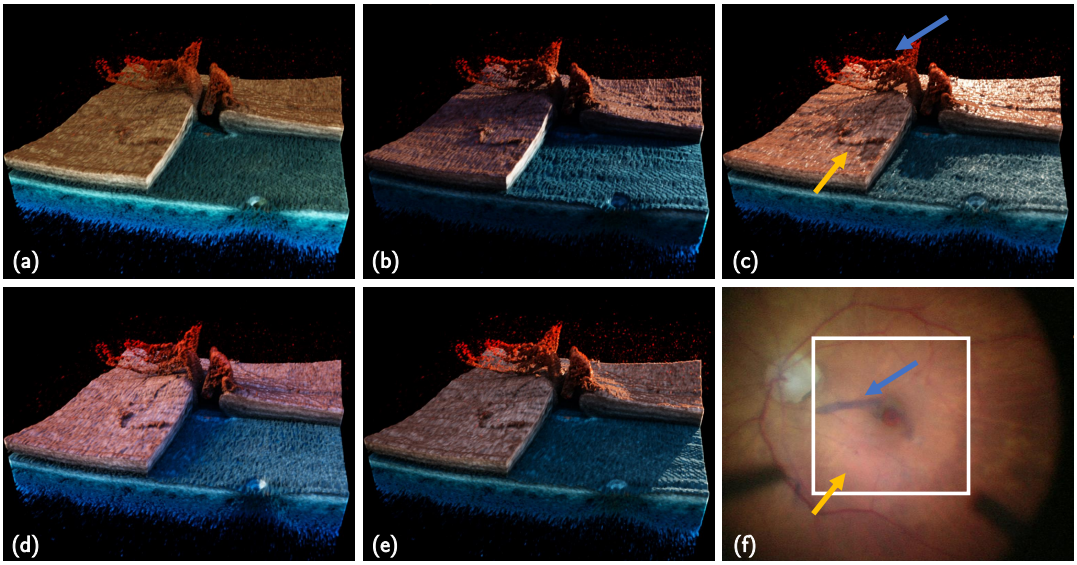


Fig. 8.4. Lighting variation study using the same transfer function and surface shading properties between all images **(a-e)**. The OCT volume shows a macular foramen after ERM peeling with peeled-away parts of the ERM covering the hole in the center (blue arrow) and visible peeling boundary (yellow arrow). **(f)** Corresponding microscope image with OCT imaging region.

in a similar spirit as Bleicher et al. [18]. We also performed qualitative comparisons of different illumination approaches (c.f. Figure 8.4), taking advantage of the flexible illumination setups possible with the software.

From these experiments we draw the following insights:

- shallow light source positioning can make a great difference in enhancing surface detail, but needs to be manually adapted to the retinal surface slope to work well (Figure 8.4(b))
- strongly colored light sources can interfere with the the color mapping and should be avoided (Figure 8.4(d))
- ambient light is essential to fill out shadowed regions (Figure 8.4(a))
- lighting from the back (Figure 8.4(c), (e)) can elicit specular highlights and and emphasize structures extruding from the surface but creates unfavorable shadows

After discussions with our experts regarding these concepts, we concluded that for an intraoperative visualization, a single, white light source placed above and to the side of the camera is sufficient when combined with some ambient lighting. Further experiments with emissive elements, for example the specific layers or the instrument, yielded visually interesting results, but were deemed of limited practical utility during a surgery.

These insights gained from the experiments with a high-quality, physically based renderer were then transferred to the development of a renderer based on volume ray casting, which is more amenable to real-time applications than the probabilistic light simulations used in prototyping.

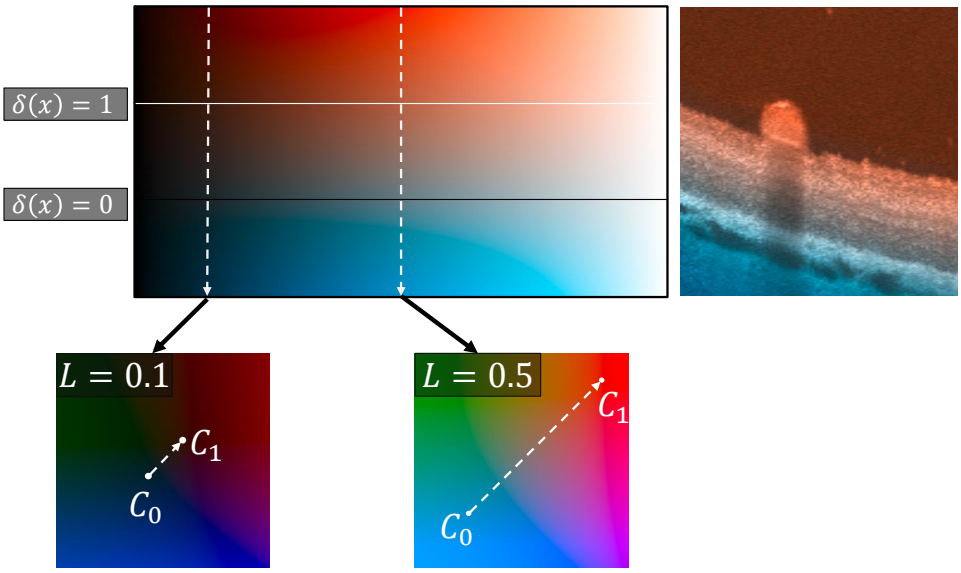


Fig. 8.5. Resulting color map for $C_0(L) = (L, -0.5, -0.5)$ and $C_1(L) = (L, 0.75, 0.75)$: $I(p)$ increases towards the right, Y axis shows changing values of $\delta(p)$. The effect of the scaling function $\gamma(I)$ is equivalent to moving the two points closer to $(0, 0)$ in the a^*b^* plane for intensity I close to 0 or 1.

8.3.2 Perceptually Linear Depth-Encoding Color Maps

Perception of depth differences is a major concern for both use cases: In ERM peeling, subtle surface irregularities are important to understand while for subretinal injection, understanding the precise needle penetration depth and distance from the target layer is crucial.

We designed a color mapping function that can map two values, axial position and OCT intensity, to a color value. To ensure good visual separation of these two parameters, we map them to different perceptual parameters: Luminance and chrominance. This is important to allow surgeons to retain structural perception of the OCT intensity while adding additional information about depth, creating more visual contrast between different positions.

We assume the existence of an RPE layer map $\mathrm{rpe} : p \mapsto [0, 1]$ giving the axial position of the RPE layer corresponding to the A-Scan containing position p. To ensure that linear changes in the parameters are also perceived linearly, we make use of the $\mathrm{L^*a^*b^*}$ color space, which is designed to be perceptually linear [206]. To map each voxel p to a color, we first define a *depth predicate* $\delta(p) \in \mathbb{R}$ which encodes the depth of a voxel relative to the RPE reference layer as $\delta(p) = (p_x - \mathrm{rpe}(p)) / d_{\mathrm{rpe}}$ where d_{rpe} is a normalization factor which is chosen to match the average thickness of the whole retinal tissue. To limit the colorization effect to differences close to the retinal tissue, we define a normalized version as $\delta^*(p) \in [-1, 2]$ as $\delta^*(p) = \mathrm{clamp}_{[0,1]}((\delta(p) + 1)/3)$. With this normalized predicate, we then define the color mapping function as an interpolation between two points $C_0 = (I(p), a_0, b_0), C_1 = (I(p), a_1, b_1)$ on the plane $L = I(p)$, as visualized in Figure 8.5:

$$C_{\mathrm{L^*a^*b^*}}(I, \delta^*) = \gamma(I) \cdot (\delta^* \cdot C_1 + (1 - \delta^*) \cdot C_0) \tag{8.14}$$

We introduce the scaling factor $\gamma(I) = 1 - (I - 0.5)^2$ which is a parabola with a maximum at $\gamma(0.5) = 1$ and $\gamma(0) = \gamma(1) = 0$. This reduces the saturation of the colors for very high and very low image intensity values $I(p)$ and keeps the color values within the visible color gamut. With this framework, we can define a class of color maps on the a^*b^* plane defined by the two points C_0, C_1. For our visualizations, we fix $C_0(L) = (L, -0.5, -0.5)$ and $C_1(L) = (L, 0.75, 0.75)$ as this gives us a color gradient from blue to red (see Figure 8.5).

Note: Arbitrary Color Maps

Parametrizing the curve as a line between two points is an easy way to define color maps, yet the process above can be generalized to arbitrary colormaps: any curve with *constant velocity* with respect to δ^* will provide similar perceptual linearity. This means that for example linearized splines could be used to parametrize arbitrary curves. To avoid the computational cost of evaluating a more complex curve on every transfer function lookup, the color map $C_{\mathrm{L^*a^*b^*}}(I, \delta^*)$ can be sampled into a 2D texture.

8.3.3 Layer-aware DVR

We have designed a volume rendering method that provides enhanced perception of superficial features and is based on the standard GPU volume raycasting pipeline [114]. We combine our color mapping with a linear opacity transfer function. Because our depth predicate δ is based on the distance to the closest RPE position, global curvature does not have an effect on our colorization. We avoid gradient-based shading by using shadow rays as discussed in [175]. This straightforward way to implement volume shadows casts a shadow ray from every sample towards the light direction, sampling the intensities to accumulate a shadowing factor. This not only avoids dependency on a noisy gradient but additionally creates more visual fidelity by adding shadows, which has been shown to improve structural perception [41]. We also test a variant where the number of steps N_s a shadow ray takes is fixed. This not only limits the performance impact on the rendering but also shortens the distance across which shadows are cast. Contrary to Viehland et al. [211], we do not use any elaborate opacity boosting but apply a color transfer function computed on the fly for every raymarch sample. Our opacity transfer function $\alpha(I)$ is a piece-wise linear transfer function with range $[I_{min}, I_{max}]$ based on intensity only. We compute the final RGB color and opacity for a sample position as

$$C(p) = \left[s(p) \cdot \mathrm{RGB}\left(C_{\mathrm{L^*a^*b^*}}\left(I(p), \delta^*(p)\right)\right), \alpha\left(I\left(p\right)\right)\right] \tag{8.15}$$

where $RGB(C)$ is an $\mathrm{L^*a^*b^*}$ to RGB color space conversion and $s(p)$ is the shadow factor (or *occlusion*) determined via casting N_s shadow steps:

$$s(p) = \prod_{0 \cdots N_s}^{i} \left(1 - \alpha\left(I\left(p + i \cdot L(p)\right)\right)\right) \tag{8.16}$$

with $L(p) = |p_{Light} - p|$ and p_{Light} being the virtual light source position.

8.4 Layer-Adjusted MIP Projection

A B-Scan aligned with the surgical needle could provide useful distance information, however imaging such a B-Scan consistently is not possible without tracking the needle to compensate for movement. Any misalignment of such a B-Scan would cause the perceived needle distance to be slightly off, leading to potentially dangerous misjudgment. Additionally, due to the needle shadow, the retinal structure below the needle would not be visible. While maximum intensity projections have shown good visualization as the maximal intensity corresponds to important pixels such as the needle or the RPE layer, these objects suffer from poor visibility when applied to an iOCT volume directly: Due to the natural curvature of the retinal surface, the target layer is not easily recognizable. While it is technically possible to constrain the MIP projection to a small area around the needle where the curvature is low enough, this would require prior knowledge or assumptions of the needle position within the volume (for example using the tracking algorithm provided in section 7). Instead, we propose a *layer-adjusted MIP* (LA-MIP) which instead of projecting along straight lines through the volume, projects along lines that are parallel to the reference RPE layer. Effectively, this produces an image where every pixel corresponds to the projection along a curve that has a constant axial distance to the reference layer. Figure 8.6 illustrates the basic idea.

Note: Coordinate System

The coordinate system chosen here is aligned with the natural memory layout of OCT data, which typically linearizes the data of A-scans in contiguous memory along the X axis, concatenates subsequent A-scans along the Y axis to form B-scans and stacks multiple B-scans along the Z-axis.

To compute an output pixel color for a screen space pixel at normalized position $P = (P_x, P_y) \in [0,1] \times [0,1]$, we set up the projection start and direction as $p_0 = (P_x, P_y, 0), d = (0,0,1)$. The reference layer distance for one ray is then $\delta_{ref}(P) = p_y - \mathrm{rpe}(p_0 + 0.5d)$. When marching through the volume, the original sampling position is $p_i = p_0 + \frac{i}{N-1} \cdot d$ for a sample with index i out of N total samples. We adapt the sampled X position to have the same distance to the sample's RPE layer height:

$$p_i^*(P) = \begin{pmatrix} \mathrm{rpe}(p_i) + \delta_{ref}(P) \\ p_{i,y} \\ p_{i,z} \end{pmatrix} \tag{8.17}$$

To compose the final color, we apply our distance encoding transfer function using the depth predicate from the reference layer:

$$C_{\text{LAMIP}}(P) = C_{\mathrm{L^*a^*b^*}}\left(\max_{i \in [0 \ldots N-1]} \{p_i^*(P)\}, \delta^*(p_0 + 0.5 \cdot d) \right) \tag{8.18}$$

Intuition: Layer-Adjusted Projection

An alternative way to think about the method is in terms of flattening: A traditional MIP with side-on viewing only works well if there is no retinal curvature along the projected direction. With the layer segmentation, we can shift every A-scan up or down to flatten the layer geometry out. Equation 8.17 performs this flattening on the fly in the viewing direction, while keeping the curvature perpendicular to the viewing direction, which does not interfere with the visualization.
It is interesting to note that this can be performed with arbitrary projections, for example the average projection used for Digitally Reconstructed Radiographs. Applying this method to other modalities and finding suitable projections could be an interesting area for future research.

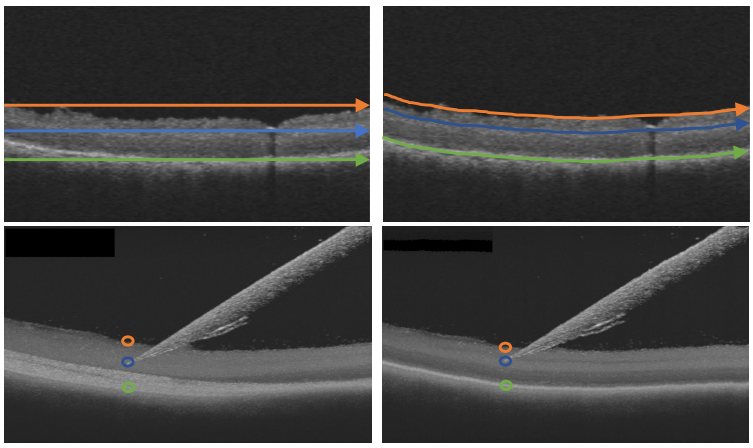


Fig. 8.6. Schematic view of our layer-adjusted MIP projection: Instead of projecting along straight, axis-aligned rays (left), we propose a projection along curved lines adjusted to the profile of the reference layer (right). This leads to better visualization of the actual distance between instrument and target layer.

We generate a projection in the y direction analogously by permuting the coordinates accordingly and combine these two views with an en face projection (see Figure 8.7), where every pixel is colored according to our color map using the maximal intensity and its respective location for colorization. For presentation, we combine these two views with a special enface projection (see Figure 8.7), where every pixel is colored according to our color map using the maximal intensity and its respective location for colorization.

8.5 Case Studies

To provide the reference RPE layer, we use the segmentation algorithm of the Cirrus HD (Carl Zeiss AG, Germany) diagnostic OCT device and import the results for our iOCT volumes into our rendering framework.

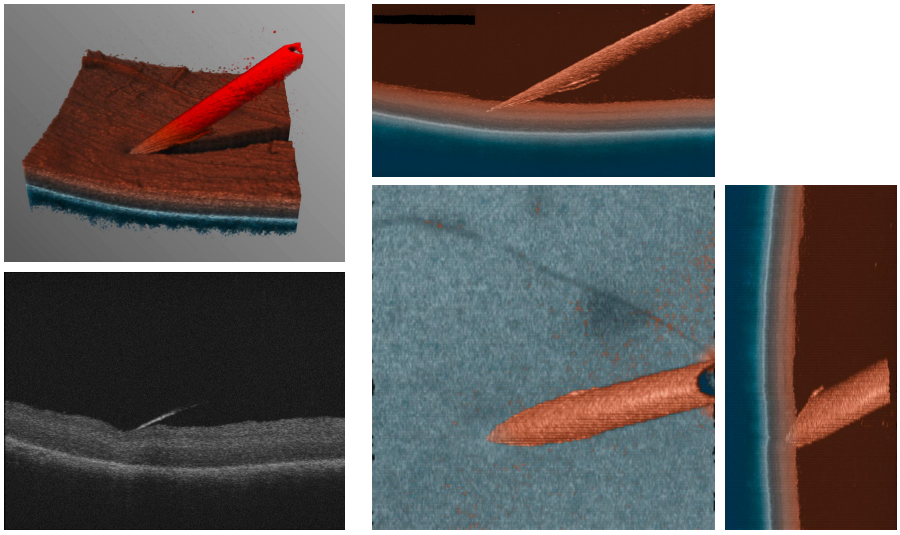


Fig. 8.7. Needle touching the retinal surface. From DVR it is not directly apparent whether the needle has already perforated while B-Scan inspection suffers from misalignment, making the information unreliable. Our final composed image of en face and LA-MIP views show clearly that the needle is below the surface, but far enough from the RPE layer.

Note: Layer Segmentation

A proprietary segmentation was used here only because it was readily available with the system software stack we used. There are many published retinal layer segmentation algorithms readily available (e.g. [115, 159, 178, 192, 207], to list a few). Many will likely provide similar or better results than the method used here.

To reduce the speckle noise apparent in our iOCT cubes while keeping small structures, we apply a 3×3 median filter on every B-Scan but do not perform the volumetric Gaussian smoothing suggested in [211]. We implemented both our visualizations using C++ and OpenGL as fragment shader programs. The shadowing factor is computed per sample on the fly without caching. For the limited length shadows, we set $N_s = 20$ to provide real time performance with small local shadows. For the full shadowing, the shadow rays are terminated a maximum of $N_s = 200$. For simplicity of implementation, we do not employ acceleration structures except for early ray termination when ray opacity $\alpha > 0.975$. For our visual comparisons, we also implemented the raymarching of [211] inside our own framework to ensure our benchmarking is comparable. Visualization parameters and transfer function were manually adapted to achieve good visibility for our OCT characteristics. All OCT scans used were acquired from a Carl Zeiss Lumera 700 operating microscope with a RESCAN 700 integrated OCT [248]. For each OCT data set we manually selected a transfer function intensity range I_{min}, I_{max} that was used for all renderings of the same data set. All our tests were performed on a machine with an Intel i7 8700K CPU and an NVidia Titan Xp GPU. We measured an average total of 1.6s for segmentation and 0.3s for volume median filtering. We benchmarked the raymarching methods by measuring average frame time for rendering a 512×512×128 volume at a resolution of 1024×1024 during one full rotation (360 frames) of the volume around the Z axis. The baseline methods [18, 211] achieve $4.6\,\mathrm{ms}$ and $5.6\,\mathrm{ms}$. Our colorized

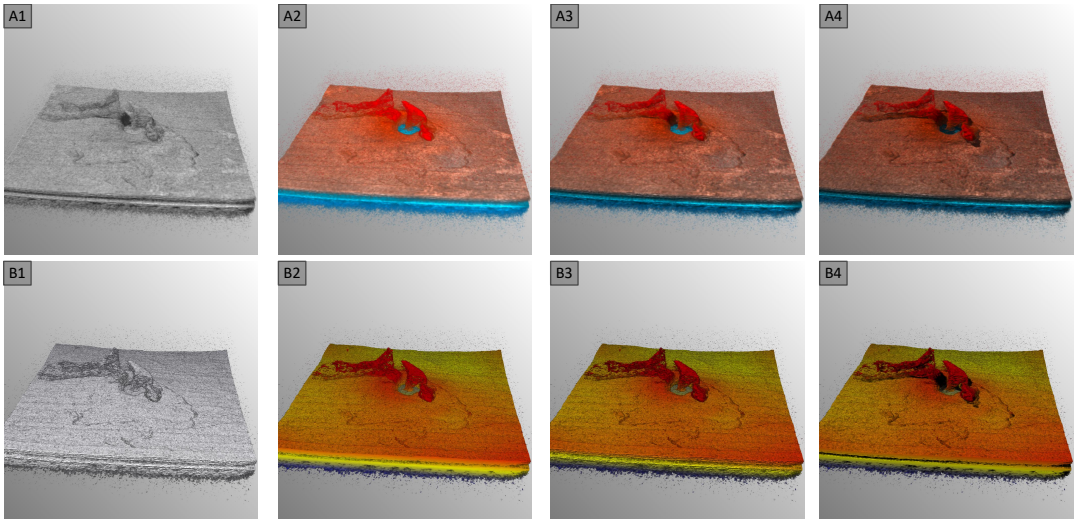


Fig. 8.8. Retinal flap positioning to close the hole with different shading options. **A1:** Local Shadow Rays only, **A2:** Depth-based Colorization only, **A3:** Limited Shadow Rays, **A4:** Full Shadow Rays **B1:** Shading and opacity enhancement [211], **B2:** Axial colorization and opacity enhancement (no Phong shading), **B3:** Visualization as in [18]. **B4:** Same as **B3** but with shadow rays instead of Phong shading.

raymarching requires $4.5\,\mathrm{ms}$ without any shading, $8.6\,\mathrm{ms}$ for $N_s = 20$ shadow steps and $31.1\,\mathrm{ms}$ for $N_s = 200$. The composed LA-MIP projection view consisting of all three projections is generated in $1.1\,\mathrm{ms}$.

Example Cases

We have selected several iOCT volumes to highlight the results when applying the different variants of iOCT visualization.

Figure 8.8 shows a case after ERM peeling. The macular hole can be seen at the center of the images. The peeled ERM is positioned above the macular hole as an *inverted flap*, which has been shown to improve healing. 3D iOCT is especially helpful in this case to check whether its placement over the hole is correct. The different shading options applied here illustrate that without any shading or with gradient-based shading, perception of the relative positioning is challenging. The variants using shadow rays produce a softer appearance and the additional shadows help understanding the 3D relationship. Our DVR method is able to provide better visualization of smaller superficial structure when compared to the baseline method.

Figure 8.1 demonstrates a needle at the ideal positioning for injection in an ex-vivo porcine eye. With our composite view of an en face view and two LA-MIP views, the needle position can be easily assessed in all three directions. Figure 8.7 exemplifies a situation where the surgical instrument touches the retinal surface. From the DVR rendered view, deformations of the surface can be monitored while the LA-MIP views allow for a better assessment of the needle tip position.

Expert Feedback

We have presented several cases of iOCT acquired during routine epiretinal surgery to two expert ophthalmic surgeons with clinical experience of 27 and 12 years, respectively. Both

Parameter	Values	Cases
Rotation Speed [deg/s]	36, 45, 60, 72, 90, 120	1
View Inclination [deg]	10, 20, 30, 40, 60, 90	2
Aspect ratio	*axially extended*, *correct*	4
Illumination	*none*, *Phong*, *limited shadow rays*, *full shadow rays*	6
Colorization	*Transfer Function*, *Axial Position RGB*, *Axial Position* $\mathrm{L^*a^*b^*}$, *Layer-Relative RGB*, *Layer-Relative* $\mathrm{L^*a^*b^*}$	6

Tab. 8.1. Parameters and number of cases for each parameter that were presented to our expert surgeons for feedback.

stated that they work with iOCT regularly, although none of them have used a system that shows 3D cubes intraoperatively in a clinical setting. In our interview session, we presented several variations of visual parameters (rotation speed, inclination, aspect ratio, illumination model and colorization scheme, c.f. Table 8.1) on a total of 8 different data sets in order to find sensible default values for an automated, interaction-free visualization. The interviewed surgeons suggest a rotation speed of about 45deg/s and an inclination angle of about 25deg with respect to the retinal tissue should work well for the presented cases. Comparing the shading options, our surgeons preferred shadow rays for all cases, judging it to be more natural and intuitive compared to Phong shading. Notably, in two cases with layer-relative colorization, the un-shaded option was judged to be of similar value as the other versions are significantly darkened. The two expert surgeons stated that limited shadow rays provide higher value due to their smaller occlusion radius. Colorization options were appreciated more when a depth-based colorization was applied than with axial colorization. Ideally, a unique color should map directly to each specific retinal layer, which is close to what we achieve for non-pathological cases. Nonetheless, one surgeon remarked that intensity-based transfer functions with a high number of color bands seem to be better at enhancing the high reflectivity of the RPE layer, demarcating it as a clear line which is sometimes desirable.

8.6 Real-Time Visualization of 4D SS-OCT

Applying our rendering method to SS-OCT is not straightforward due to the higher performance requirements and increased bandwidth of SS-OCT. In cooperation with the Center for Medical Physics and Biomedical Engineering at the Medical University of Vienna and supported by Carl Zeiss Meditec, we have adapted our rendering algorithm for integration with a state-of-the-art SS-OCT engine[20]. Results from the system, running at 17 volumes per second, can be seen in the screenshots in Figure 8.9.

System Architecture

The OCT imaging engine [20] is built around a MEMS-tunable VCSEL light source with a central wavelength of 1060 nm with a maximum sweep repetition rate of 1 MHz. A 12-bit PCIe digitizer card acquires the spectral samples at 4 GSPS. Using spectral splitting [71], the system is able to achieve up to 2 MHz effective A-scan rate at an axial resolution of $12.6\,\mu\mathrm{m}$

(FWHM) in tissue with an imaging depth of 4.3mm, or increased imaging depth up to 29mm at 100kHz A-scan rate. The imaging system uses spiral scanning [25] for 3D imaging to minimize scanner flybacks and strain on the galvo scanners.

We implemented a multi-GPU processing system similar to previously reported systems [111, 235]: One GPU is dedicated to performing the OCT image reconstruction fully in CUDA. This GPU also performs remapping of the spiral A-scan layout to a Cartesian grid for rendering. The rendering is implemented in CAMPVis running as a separate process on a second GPU dedicated to the rendering pipeline. We use the OpenGL-CUDA interoperability API in combination with the CUDA interprocess API to transfer each volume directly to the OpenGL context of a the rendering process. The GPU-GPU transfer is performed via NVLink to bypass CPU memory, which provides improved latency and avoids interfering with the data stream from the digitizer card on the PCIe bus. The PC system is connected to a passive 54" stereoscopic display with a native resolution of 3840×2160 pixels via HDMI.

Algorithm Adaptation

To support real-time rendering of the 4D volume data in a stereoscopic view, several adaptations were necessary:

Firstly, real-time segmentation of B-mode data at this data rate is not easily achievable with current segmentation models and is an area of active research [19, 192]. To avoid this complexity, we resort to a proxy map that is much simpler to compute: we use the centroid map, which for each A-scan at x, y computes the centroid of the image intensities I, giving a position along the A-scan that is roughly aligned with the high intensities of the retinal tissue (note that A-scans are along the first axis in I):

$$P_{\text{centr}}(x, y) = \sum_a aI(a, x, y) / \sum_a I(a, x, y) \tag{8.19}$$

We compute this efficiently via a compute shader on the GPU as part of the rendering pipeline. This centroid map is good enough to align the layer-aware colorization used in our 3D rendering. The LA-MIP on the other hand requires a precise anatomical reference layer to work well, so we cannot use this visualization for the 4D intraoperative rendering in this system. Instead, we resort to the colored enface MIP projection introduced in section 8.4 in combination with two perpendicular MPR planes for cross-sectional details.

Note: Real-Time OCT Segmentation

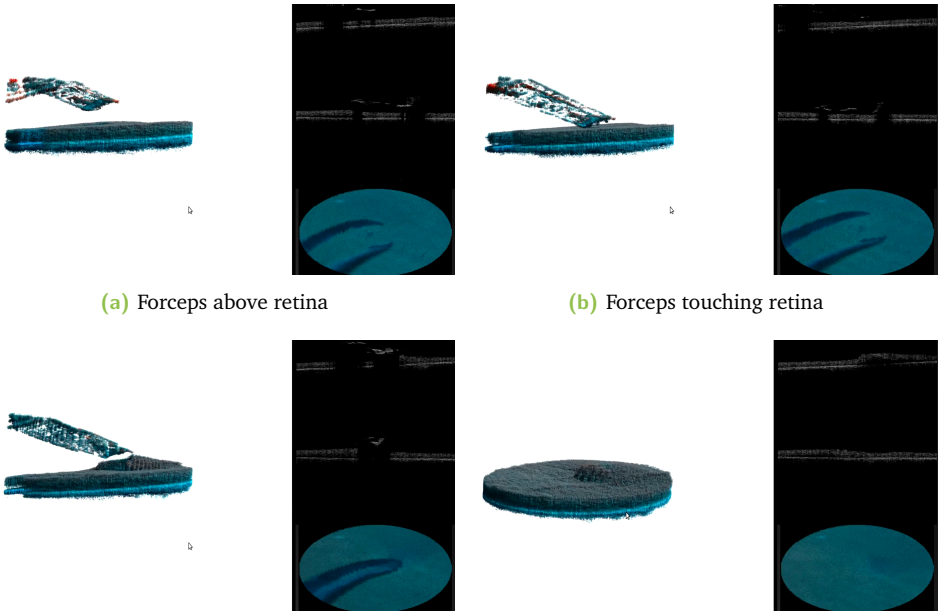
Recent work by Borkovkina et al.[19] seems to put online segmentation within reach. They show that with an optimized UNet architecture, the use of Tensor Cores[5] and 8 bit precision math, real-time segmentation of OCT is possible. They perform real-time OCT reconstruction and inference on a single NVidia RTX 2080 Ti GPU and report 3.5ms for inference of B-Scans consisting of 400 A-scans, which corresponds to a segmentation rate of approximately 114000 A-scans/s. While this is still off by more than an order of magnitude from our maximum A-scan rate, more recent hardware and network architectures in combination with some downsampling could put full real-time segmentation within reach.

Monoscopic rendering			
	Volume Resolution	**Display Resolution**	**Timing**
Basic Raycasting	2048×400×400	3840×2160	55.1 ms
Phong Shading	2048×400×400	3840×2160	63.5 ms
Shadow Rays ($N_s = 10$)	2048×400×400	3840×2160	111 ms
Shadow Rays ($N_s = 20$)	2048×400×400	3840×2160	144 ms
Stereoscopic rendering (horizontally interlaced)			
Basic Raycasting	2048×400×400	3840×2160	83.1 ms
Phong Shading	2048×400×400	3840×2160	95.4 ms
Shadow Rays ($N_s = 10$)	2048×400×400	3840×2160	151 ms
Shadow Rays ($N_s = 10$)	2048×400×400	2688×1512	95.4 ms
Shadow Rays ($N_s = 10$)	1024×400×400	3840×2160	68.1 ms
Shadow Rays ($N_s = 10$)	1024×400×400	2688×1512	42.1 ms

Tab. 8.2. Performance analysis of the DVR rendering, performed with an NVidia 2080 Ti, with basic raycasting (pure Emission/Absorption) and gradient-based Phong shading for comparison.

Unfortunately, at the high temporal and spatial resolution the imaging system provides, rendering the DVR visualization itself also becomes a bottleneck. We apply the following modifications to reduce the rendering time: The number of shadow ray steps was reduced from $N_s = 20$ to $N_s = 10$. This practically removes any shadowing, however it still retains the soft shading that is important for surface perception. Furthermore, given the geometry of the spiral scanning pattern, we know that the 3D volume of the remapped OCT data only contains data within a defined cylinder. We use this known cylindrical geometry as our proxy geometry during the computation of the entry and exit points for raymarching [114] as a simple but effective way to perform empty space skipping. Finally, we performed a performance analysis of the rendering algorithms with respect to input volume resolution and rendering resolution (c.f. Table 8.2). In response to the analysis, we introduced variable downsampling factors for the input volume (implemented via trilinear interpolation in a compute shader) and the output rendering resolution. These can be dynamically adjusted to optimize performance depending on the current imaging parameters of the OCT engine. From our experience, a resolution scale as low as 0.7 (which results in 51% reduction in rendered pixels) is still barely noticeable at the given display resolution when viewed from the recommended minimum of 2m distance from the display.

For volume acquisitions, our OCT engine operates at an effective A-scan rate of 800kHz - 1MHz. However, to achieve this high scan rate, spectral splitting with a factor of 2 is employed which effectively halves the axial resolution. Therefore, each A-Scan has a resolution of 1024 in depth. The spiral scan geometry is remapped to a cartesian grid with a similar spacing which leads to a resulting volume resolution of 1024×293×293. With this volume resolution, the tweaked randering algorithm can achieve 17Hz at full resolution (interlaced 3840×2160), thus the viewport resolution downscaling methods are only necessary for higher volume rates or volume resolutions.



(a) Forceps above retina

(b) Forceps touching retina

(c) Forceps pulling on retina

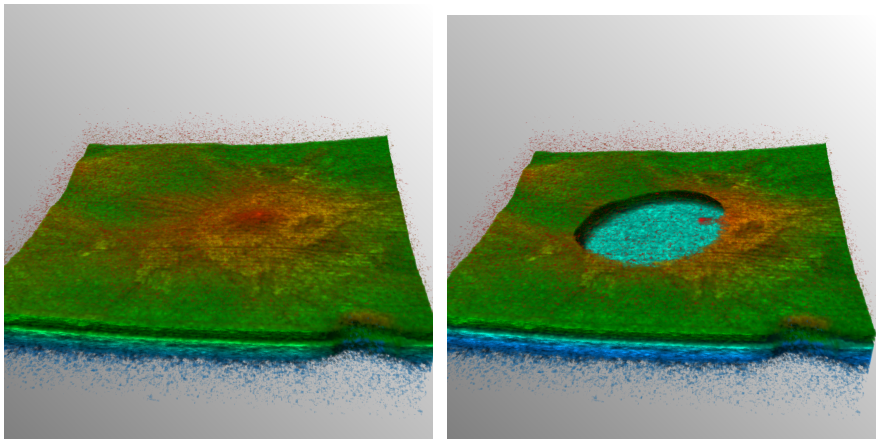
(d) Retinal detachment

Image source: 2021 Courtesy P. Matten, with permission

Fig. 8.9. Image sequence from the layer-aware rendering adapted to 4D OCT, running at 4k stereoscopic rendering resolution (with horizontal interlacing) on a system with two NVidia Titan RTX connected via NVLink. Sequence shows a forceps inducing retinal detachment in an ex-vivo porcine eye, imaged at 17 volumes/s.

8.7 Conclusion

The visualization concepts we have proposed both rely on a segmentation of an anatomical layer that can be used as a reference. In our experiments we used an existing implementation that was not optimized for iOCT volumes and thus sometimes exhibited areas of faulty segmentation, especially in the presence of instruments. An incorrect segmentation map can strongly influence the quality of the visualization, potentially leading to distorted projections and confusing colorization. With the ongoing advances in deep learning, more robust segmentation methods are readily available and in combination with advances in compute hardware are approaching capabilities of real-time segmentation. We have shown two visualization methods for advanced intraoperative feedback, based on the idea of distance-based colorization with respect to an anatomical layer. To support this, we developed a color mapping that encodes both intensity and depth in a perceptually linear way to ensure good perception of both dimensions separately. In an extension of this work, we have integrated our rendering with a state of the art 4D SS-OCT engine and describe the adaptations and concessions necessary to allow for real-time rendering of volume rates up to 17 Hz.



(a) Alternative colormap (b) Ellipsoidal contextual cutout

Fig. 8.10. Possible future directions for research are given by exploring a different colormaps or b further developing the context-sensitive cutouts (section 8.3.1) to intraoperative data.

Intuition: Future work

Our experts have suggested that with the presented color map, superficial structures do not get much contrast. With the color mapping framework in section 8.3.2, it is straightforward to explore other color maps. An example can be seen in Figure 8.10a. With better availability of high-quality automatic layer segmentation, revisiting the concept for contextual cutouts (c.f. Figure 8.10b) could be another way to provide more expressive and focused visualizations.

In this chapter, the focus within our proposed visualization-guided robotic system was primarily on the requirement for more advanced visualization (**III.**) based on an improved semantic understanding through the known location of the layered anatomy (**IV.**). We have already shown how our 3D rendering can be adapted to the higher performance requirements of stereoscopic displays, albeit with some concessions in quality. VR visualization sets even tighter bounds for rendering performance, and it seems that with the higher difference between required target framerate (>60FPS) and volume rate ($\sim 20Hz$), precomputed illumination volumes [199] provide a way to mitigate the cost of our current per-sample shadow ray marching. Nonetheless, even if intraoperative layer segmentation is only available at a low update rate we believe that both our layer-aware DVR method and the LA-MIP will be excellent tools in such an environment.

9 Processing-Aware Real-time volume rendering

Contents

9.1 Introduction

The methodology we have presented so far in this dissertation have been focused mainly on leveraging the data from previous-generation SD OCT systems. Technological advances such as the application of swept-source lasers [74], spectral splitting [71] and linear velocity spiral scanning [25] have made continuous 4D intraoperative OCT imaging feasible at high resolution and volume rates. State of the art systems are able to sample with volume rates of up to 24.2Hz at a resolution of 330×330×595 voxels [20, 111]. Ehlers et al. [55] have shown the effectiveness of 2D iOCT in clinical practice in a large scale study, however clinical studies on 4D OCT do not yet exist. Still, 4D OCT is poised ideally to improve spatial visualization

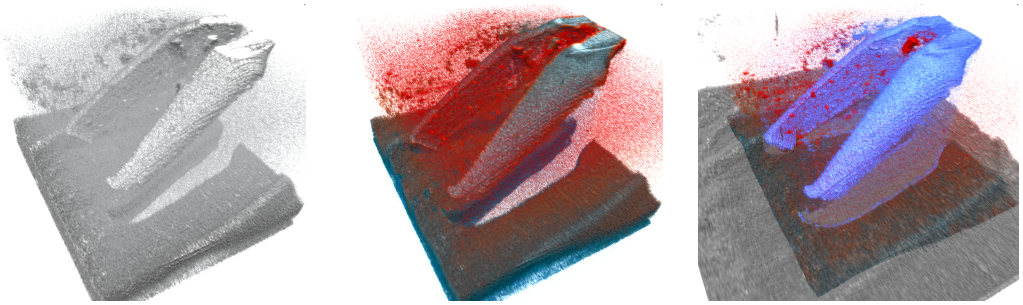


Fig. 9.1. Challenges and our solution to 4D iOCT rendering: Straightforward DVR *(left)* is not sufficient for good visualization and advanced color schemes [215] *(middle)* do not cope well with noise and artifacts. Our novel method *(right)* reduces noise and artifacts and at the same time extends the apparent field of view. All images were rendered with the same opacity transfer function.

during complex and precise maneuvers. Potential applications include robotics [241] or even retinal surgery under exclusive OCT guidance [111] which could greatly reduce the risk of phototoxic effects associated with the endoilluminators presently used for traditional fundus view surgery. With data rates surpassing 6GB/s, 4D iOCT is a uniquely challenging modality for advanced processing in an intraoperative setting. OCT-only surgery is currently challenged also by higher noise levels of intraoperative OCT compared to diagnostic OCT, instrument shadows hiding relevant anatomical structures and a limited field of view.

In this work, we propose a real-time processing and rendering pipeline that combines intermittent wide field-of-view scans with real-time imaging of a smaller focus field of view. This combination not only provides a larger effective field of view but also allows us to enhance instrument visibility and reduce noise by accumulating information temporally.

Key Contributions

- We introduce a fast registration and compositing method that temporally combines the static parts (e.g. the retinal surface) within the OCT imaging region
- This is supported by a learning-based instrument segmentation in 2D projection images of the OCT volume
- We further propose a processing-aware rendering method that optimizes tissue visibility by inpainting shadowed areas, ensuring interpretability of the view through sensible color mapping

This chapter has been published previously as [219].

9.2 Related Work

To improve quality in diagnostic OCT imaging, a typical approach is to perform repeated scanning of the same location and average the signal to improve SNR [10, 201], either by averaging the raw signal or the reconstructed B-scans. Because this approach would reduce volume rates to impractical speeds for intraoperative use, an alternative solution is to average subsequent volumes, yet this requires some form of motion compensation to avoid "smearing" artifacts. Previous works on OCT motion compensation are focused on motion during acquisition of a single volume [112] or on registering diagnostic OCT volumes acquired at different visits for progression analysis. 3D SIFT features [155] have been used for full 3D registration of OCT volumes in diagnostic scenarios. Transverse alignment of 2D projections has been shown based on 2D vessel segmentation by ICP [156] as well as matching SURF keypoints [160]. Gan et al. [67] use SIFT feature based matching on 2D en face images and then matched the scan edges at the overlapping regions. However, high computation times of these methods make them not suitable for our real-time use case. Additionally, due to the

relatively small field of view of typical intraoperative OCT and a potential absence of vessels and distinguishable structures, these feature based algorithms can result in bad registration performance. To the best of our knowledge, there is no published research on solving this problem with the additional constraints of real-time processing and a potentially moving surgical instrument confounding the registration.

Metallic instruments pose an additional challenge to iOCT visualization, causing total loss of signal below the instrument due to the metal blocking the light. This shadows relevant retinal tissue below the instrument, causing "holes" in the 3D rendering. Special OCT-compatible instruments have been proposed [56], but are not yet translated to clinical practice. Effective visualization of iOCT volumes is not trivial due to the aforementioned problems, which is why straightforward methods like direct volume rendering (DVR) with intensity-based transfer functions perform badly (c.f. Figure 9.1, *left*). Viehland et al. [211] introduced a DVR method for OCT that was extended by Draelos et al. [18] to add colorization along the axial direction. Both however rely on spatial filtering with Gaussian and median filters to reduce noise prior to rendering. Layer-aware volume rendering (c.f. section 8) uses a layer segmentation to anchor a color map to improve perception (c.f. Figure 9.1, *middle*). Our aim in this new work is to show the feasibility of advanced real-time processing and visualization for 4D iOCT. Our novel real-time processing and rendering method combines temporal information and mitigates instrument opacity (c.f. Figure 9.1, *right*).

9.3 Methods

Our visualization was built around the two goals of reducing imaging noise over time and using data from previous frames to fill in the gaps produced by instrument shadowing. The approach is based on the notion that with 4D iOCT, retinal tissue is imaged many times while remaining relatively unchanged. Therefore, we discriminate the retinal tissue to align it over time and use the aggregated information to improve our visualization. Due to the high data rates and time constraints we rely on processing 2D projection images instead of 3D volumes.

Our novel method consists of several steps (see Figure 9.2): To initialize our algorithm, a larger reference volume is acquired to prime our compounding volume I_c and establish a reference. For each newly acquired volume I_a and once for the reference volume, we generate a set of 2D projection images and use a learning-based approach to find a tissue mask M_t by segmenting instruments in 2D. The masked projection images are then used to register the I_a to I_c. The registration together with M_t allows us to maintain an updated representation of the static retinal tissue while excluding the instrument and shadows. Finally, our enhanced rendering selectively combines the two volumes to provide an responsive visualization. We leverage the instrument mask and projection images to combine I_c and I_a on the fly during rendering and adapt the rendering parameters, encoding the reliability of the information through color.

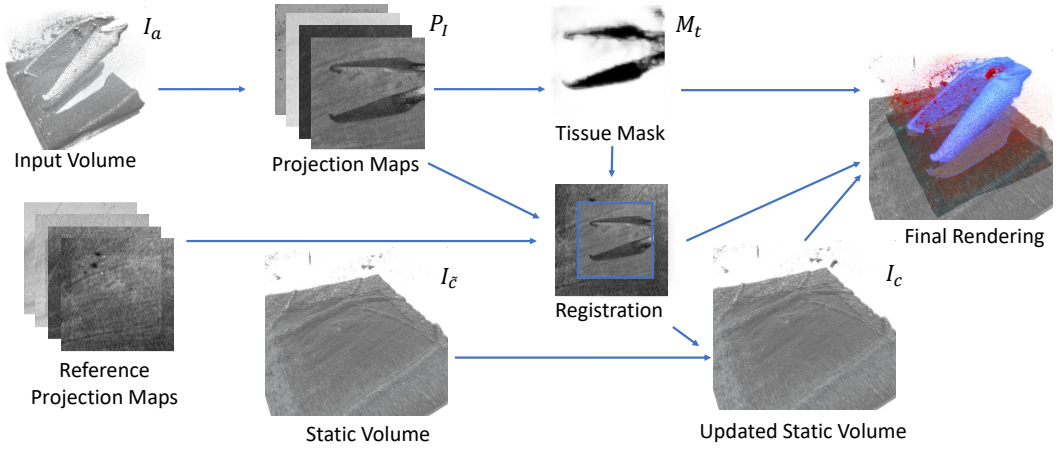


Fig. 9.2. Our processing pipeline for a single new acquired volume. Reference projection maps are generated from an overview scan at acquisition start.

Intuition: Tracking, Reconstruction and SLAM

Originally we experimented with a variant of this pipeline that does not rely on scanning a dedicated reference volume, instead initializing from the first volume and continuously updating the reference projection maps. This was inspired by the SLAM (Simultaneous Localization and Mapping) approaches in computer vision which can build up a 3D representation of the environment over time from a camera stream based on multi-view geometry. Unfortunately, when transferring this approach to OCT, we encountered a famliliar problem: continuously updating the projection maps will accumulate tracking errors will cause drift and eventually diverge completely. The reference projection maps provide a fixed reference, thus avoiding accumulation of errors.

9.3.1 Axial Projection Images

Axial projection images are a class of 2D images that we generate from projecting every A-scan in the volume to a single corresponding value. Averaging projections have been used extensively in the past [67, 156, 160] as so-called *Enface projection* images. We generalize them to new functions which create 2D feature maps $P(x, y)$ with different characteristics that we use for alignment and instrument segmentation:

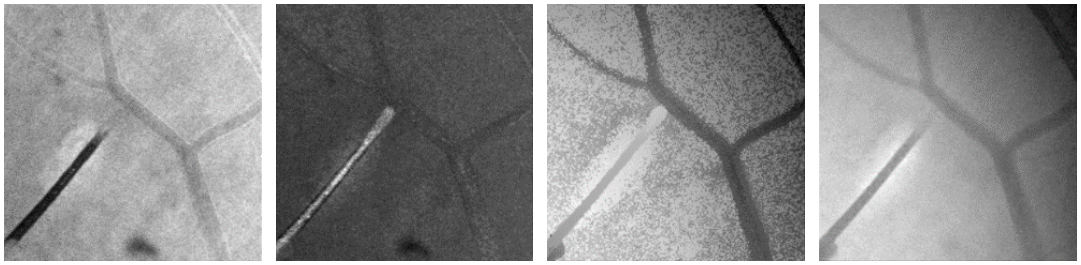
$$P_{\text{avg}}(x, y) = \frac{1}{N} \sum_z I(x, y, z), \tag{9.20a}$$

$$P_{\text{max}}(x, y) = \max_z I(x, y, z), \tag{9.20b}$$

$$P_{\text{argmax}}(x, y) = \operatorname*{argmax}_z I(x, y, z), \tag{9.20c}$$

$$P_{\text{centr}}(x, y) = \sum_z zI(x, y, z) / \sum_z I(x, y, z) \tag{9.20d}$$

where N is the number of samples in one A-scan.



(a) P_{avg} (b) P_{max} (c) P_{argmax} (d) P_{centr}

Fig. 9.3. Axial projection maps generated from the 3D volume by accumulating the values along each A-Scan with different functions. a and b encode intensity features (brighter feature = brighter voxel intensity) while c and d encode positional information (brighter feature = deeper in the scan).

Note: Coordinate System

As the methods in this chapter are primarily based on projective images that are aligned with the fundus view, it is convenient to define the volume axes accordingly. Therefore, volume coordinate system in this chapter is defined such that the x and y axes are aligned with the x and y axes of a fundus image while the z axis points into the retina and thus along an A-scan.

P_{avg} and P_{max} correspond to the familiar average and maximum intensity projections. P_{argmax} yields the position of the brightest voxel along the A-scan, usually corresponding to the instrument or the RPE layer. P_{centr} is the intensity-weighted mean position and gives a more general estimate of where the main reflective structure in the A-scan is located. Note that *Average* and *Maximum* encode intensity features while *Argmax* and *Centroid* provide positional information. All four features can be computed computed simultaneously in a single sweep over the A-Scan voxels. Figure 9.3 shows how each of these projection maps encodes different information.

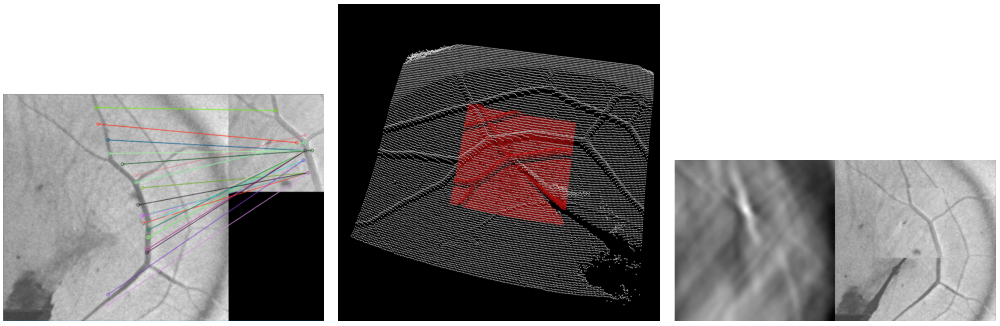
9.3.2 Learning-based instrument segmentation

A tissue mask M_t is obtained from the 2D projections by training a simple and lightweight segmentation model. We use the pytorch-based fast.ai library to train a UNet variant[1] that uses a ResNet18-based encoder that was pretrained on ImageNet connected to a decoder with skip connections to perform subsequent upsampling. All four available projection images are provided as input features to the network to maximize accuracy.

9.3.3 Registration

To combine the sequence of volumes, it is required to know for every voxel the corresponding location in the compounded volume. During a surgery, the imaged region changes over time

[1] https://docs.fast.ai/vision.models.unet.html



(a) Feature Matching with SURF+FLANN (b) Point Cloud Alignment (c) Template Matching

Fig. 9.4. Experiments with different registration alternatives: **(a)** Feature matching often cannot find reliable feature matches to compute the homography. **(b)** Point cloud alignment would only match a very coarse relative transform. **(c)** Template matching typically works if the template is not distorted.

due to systematic changes of the galvo offsets (which are known and can be calibrated for), but also due to relative movement of the patient's eye.

Potential Registration Algorithms

In a thesis defined and supervised by the author of this dissertation [191], we evaluated potential alternatives and developed the basic registration approach, of which we present a further refined variant in the following paragraphs. Building on the parallels of our problem to traditional computer vision problems, we identified the following potential registration approaches:

1. Dense volumetric alignment via iterative optimization of similarity measures
2. Feature-based matching based on SURF [11] + FLANN [150]
3. 3D Registration of surface point clouds [180]
4. Dense 2D template matching

We discarded volumetric alignment due to the generally high computational cost which limits its real-time applicability. In preliminary experiments with the three methods (c.f. Figure 9.4), it quickly became apparent that neither the feature matching nor the point cloud-based approach would produce satisfactory results. Feature matching methods would generally fail to identify reliable keypoints or produce incorrect matches, which would distort the estimated relative pose. Point cloud matching could only align the rough surfaces, but was unable to take the smaller details like the surface vessels into account to produce a better match. 2D template matching showed the most promise, and comparison of different optimization metrics[2] revealed Normalized Cross Correlation (NCC) to yield the best matches. However, in our experiments we have seen that template matching methods in general are sensitive to deviations between template and image. This is routinely the case for us, e.g. when an instrument is visible in the volume, and in those cases the matching is often offset if an instrument is visible in the template. In our adaptation (see below), we solve this by using the tissue mask M_t to exclude pixels from matching which might confuse the alignment.

[2]We compared the six different metrics available in OpenCV: Sum of Squared Differences, Cross Correlation, Correlation Coefficient, as well as their normalized counterparts

Motion Model

With the eye anaesthetized and fixed in the eye socket, rotations generally only occur during brief, intentional adjustments of the view by the surgeon. This is normally avoided during precise manipulation in which our OCT image guidance is used: the surgeon actively stabilizes the eye and thus we consider one degree of freedom less than [160] for our registration. Due to our exponential averaging, only the most recent few volumes influence the displayed result. Analyzing **DS3** (see section 9.4) using offline volumetric registration confirms that axial rotation is negligible in our data set as rotation between consecutive volumes is below 0.5°. We approximate the motion as a pure translational change in the OCT coordinate system. Two corresponding positions in reference and acquired volume p_r, p_a are related by $p_r = p_a + t_a$. We further decompose the translation t_a can be decomposed into the transverse translation (t_x, t_y) (mainly caused by rotational movement of the eye around its center) and the axial translation t_z (typically caused by inadvertent pressure on the trocars pushing the eye into or out of the eye socket).

Transverse Alignment

Because the transverse alignment (t_x, t_y) is aligned with the layout of the generated projection images P_r, P_a, we can find the best alignment using template matching with normalized cross correlation (NCC) as a metric. To find the alignment, we first center-crop the projection image P_r with a cropping factor $\gamma \in [0, 1]$ that specifies the size of the cropped image as a fraction of width and height. The cropped image $\tilde{I}_1$ is then used as the template and the transverse alignment is found as:

$$(t_x, t_y) = \operatorname{argmax}\left(NCC_{\tilde{M}_t}\left(T_0, \tilde{T}_1\right)\right) \tag{9.21}$$

where $NCC_{\tilde{M}_t}$ is the normalized cross correlation image between the image and the template while considering only the pixels included in the the (cropped) tissue mask $\tilde{M}_t$.

Axial Alignment.

To find the axial alignment t_z, we use the P_{argmax} images in the region where the two images overlap based on (t_x, t_y). We compute the average distance across the overlap region O while taking into account the tissue mask M_t of the current image:

$$t_z = \sum_{p \in O} M_t(p)\left(P_{r,\text{argmax}}(p) - P_{a,\text{argmax}}(p)\right) / \sum_{p \in O} M_t(p) \tag{9.22}$$

9.3.4 Compounding

With a known transformation we can integrate the new volume I_a in our compounded representation I_c of the static scene:

$$I_c(p) = \omega(p) I_a(p + t_a) + (1 - \omega(p)) I_{\tilde{c}}(p) \tag{9.23}$$

for every voxel position p in I_c where $\omega(p) = \omega_0 M_t(p)$ and ω_0 is a integration weighting parameter used to control the amount of exponential averaging over time. The update conditionally integrates new data for A-Scans where M_t is not set while retaining data from

the previous compounded volume $I_{\tilde{c}}$ otherwise. Using the M_t in the integration weighting masks out A-scans containing the instrument while averaging the retinal tissue.

9.3.5 Rendering

To provide a well interpretable rendering that makes the the reliability of the shown data apparent, we use an adaptive colorization that extends the color we introduced in section 8.3.2. Our goal is to emphasize which parts of the volume are from the last acquired volume and visually set apart data that has been inpainted from previous data. Thus, we differentiate three semantic regions within our volumes:

(a) data that has been recently updated or is sampled directly from I_a,
(b) data that has been inpainted from the I_c with no correspondence in I_a, and
(c) areas surrounding the instrument.

When determining the sample intensity I_s for a raymarch step position p_s, we combine I_c and I_a using an instrument predicate κ_i and a recency predicate κ_r defined as:

$$\kappa_i(p) = \neg M_t(p) \wedge |p_z - P_{\text{argmax}}(p)| < d_i, \tag{9.24a}$$

$$\kappa_r(p) = M_t(p) \wedge I_a(p) > I_c(p). \tag{9.24b}$$

In A-scans that contain an instrument, we combine the intensities of both volumes with a $\max$ operation. The $\max$ operation ensures visibility of both instrument and inpainted tissue in areas of signal loss:

$$I_s(p_s) = \begin{cases} I_c(p_s) & \text{if } \kappa_s(p) \\ \max(I_c(p_s), I_a(p_s)) & \text{otherwise.} \end{cases} \tag{9.25}$$

The recency predicate therefore marks sampled locations in which I_s was influenced by the most recently acquired volume, either because it was incorporated into the compounding (Equation 9.23) or used directly (Equation 9.25). In tissue regions, we apply the layer colorization $C_{\mathrm{L^*a^*b^*}}(I, \delta^*(p))$ introduced in section 8.3.2. We compensate for the lack of a layer segmentation by instead using P_{argmax} to align the axial color mapping to the retinal structure. Furthermore, the colorization is only applied for pixels that have not been inpainted from the compounded volume, using a grayscale colormap for this data instead. For instrument areas, we use a fixed color C_i to further enhance instrument visibility and contrast to the surrounding tissue.

$$C_s(p_s) = \begin{cases} C_i & \text{if } \kappa_i(p_s) \\ (I_s, I_s, I_s) & \text{if } \neg\kappa_r(p_s) \\ C_{\mathrm{L^*a^*b^*}}(I_s, \delta^*(p)) & \text{otherwise.} \end{cases} \tag{9.26}$$

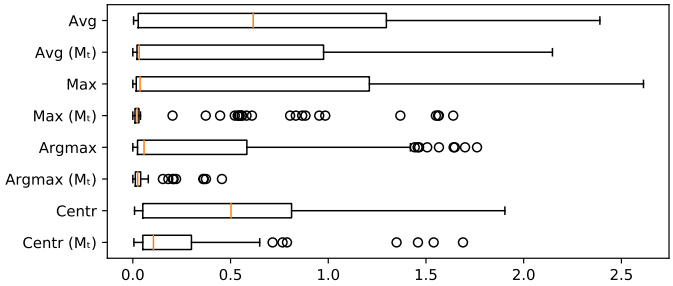


Fig. 9.5. Registration error (in mm) without using a tool mask compared to the registration error using M_t.

Projection	Avg Error	RMSE	Std Dev
Average	0.5379	0.6743	0.6516
Maximum	0.1394	0.2034	0.3329
Argmax	**0.0426**	**0.0583**	**0.0675**
Centroid	0.2128	0.2985	0.2817

Tab. 9.1. Comparison of projection images in terms of their registration performance using M_t (errors in mm).

9.4 Results

As state of the art 4D iOCT systems are not readily available, the three data sets used in this work were instead recorded with a RESCAN 700 (Carl Zeiss Meditec) iOCT system (27kHz A-Scan rate) which acquires volumes of 512×128×1024 resolution at non-interactive rates. **DS1** consists of projection images from 605 retinal OCT scans containing a range of different instruments and situations in human or porcine eyes. We generated projection images as described in section 9.3.1, resample them to 128×128px and manually labeled the instruments. **DS2** consists of $3\times3\times2.9\mathrm{mm}^3$ OCT volumes from 6 enucleated porcine eyes with removed anterior segment. On a fixated eye, a $6\times6\times2.9\mathrm{mm}^3$ reference volume and grids of 5×5 volumes were acquired with $0.25\mathrm{mm}$ offset in x/y dimensions by changing the OCT acquisition offset. In three eyes, a 23G vitreoretinal forceps was used, in the remaining three a 27G injection needle. **DS3** is a sequence of 67 volumes from an enucleated porcine eye with intact anterior segment. The first volume is acquired at $6\times6\times2.9\mathrm{mm}^3$ while the following are recorded with $3\times3\times2.9\mathrm{mm}^3$ field of view. A surgical instrument (23G retinal forceps) was moved between acquisitions to create a simulated 4D OCT sequence that shows instrument and retina motion as described in Sec. 9.3.3.

Registration

To evaluate our registration approach, we use the known relative offsets between volumes in **DS2**. We use the reference volume of each eye and register each of the 25 grid volumes to it using our approach. Figure 9.5 shows the registration error $e_r = \sqrt{d_x^2 + d_y^2}$ of the offsets d_x, d_y after registration for different projection images. The results show that using M_t dramatically improves the registration. Overall the *argmax* image performed best with an average offset error of $0.0426\,\mathrm{mm}$, RMSE of $0.0583\,\mathrm{mm}$ and median error of $0.025\,\mathrm{mm}$.

	Source	Resampled
Volume Resolution	512×128×1024	330×330×595
Input Projections	1.92	1.81
Segmentation	7.41	6.68
Registration	11.91	20.80
Compounding	7.61	8.49
Rendering	19.00	9.13
Total	**47.85**	**46.91**
FPS	**20.90**	**21.32**

Tab. 9.2. Average processing time in ms for volumes at source resolution and resampled resolution.

Instrument Segmentation

We train our segmentation model on **DS1** with a batch size of 64 and random data split of 80% training and 20% validation set. We employ a set of dataset transformations to mitigate the fact that the data set itself is relatively limited in size. Augmentations include vertical and horizontal mirroring, random rotation up to 10 degrees around the image center and random cropping and resampling to zoom up to a maximum scaling factor of 1.5. We use cross-entropy as our loss function and train with a learning rate of 0.01 and the AdamW [131] optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$). Our model converged after 95 epochs and has a validation accuracy of 99.34% and 85.33% for retina and instrument classes, respectively.

Real-time processing

Our real-time visualization system was implemented in C++ within the open-source visualization framework CAMPVis [185]. We use the OpenCL-accelerated template matching implementation of OpenCV. Furthermore, we leverage TensorRT[3] to optimize and execute our trained model on the GPU. During optimization of the model, we disable TensorRT optimizations that could potentially reduce segmentation accuracy such as mixed precision inference. Projection image generation, volume compounding and rendering leverage the GPU using OpenGL, implemented as compute and fragment shaders. We evaluated computational performance by looping **DS3** with data preloaded to the GPU to simulate GPU reconstruction as in [111]. To compare our implementation to the data rates of a state of the art SS-OCT device, we resample our input data to a matching resolution of 330×330×595. Table 9.2 shows the average results on the original and resampled data sets on our evaluation system (Intel Core i7-8700K @3.7GHz, NVidia Titan Xp). The total processing time dictates both maximum frame rate rate of our setup and minimum latency from acquisition to display. Our achieved 21.32 frames per second is close to 24.2Hz volume rate rate reported by [111].

[3] https://developer.nvidia.com/tensorrt

9.5 Conclusion

We presented the first end-to-end real-time pipeline to process and visualize 4D OCT data while mitigating instrument shadowing. Our adaptive visualization makes this processing and enhancement visible to the viewer and aids interpretation of the shown data. We demonstrated the feasibility of such a pipeline, achieving real-time processing speeds that, with minor upgrades in computational hardware or more optimized implementation, can keep up with state of the art SS-OCT systems.

Our evaluation reveals potential for future improvements for the registration approach. The implementation is based on off-the-shelf solutions for template matching and 2D image segmentation. Our matching errors are in the order of several px which leads to potentially oversmoothed results. 2D image alignment is a well-researched field in the computer vision community, and DL-based approaches to image alignment such as [128, 186] promise more robust results than our chosen template matching method. Given that the template matching used in our implementation already requires around $12 - 21\,\mathrm{ms}$, it seems feasible that a DL-based image alignment method can fit within the same time budget or even improve upon this bottleneck given the available hardware and software optimizations available for efficient inference execution. Because the focus of this work was to introduce the general concept and benefits of temporal alignment in 4D OCT, we consider the evaluation and development of more robust registration solutions as future work.

Another interesting challenge is how to mitigate larger deformations that are caused for example by tool-tissue interactions. With the current approach, these deformations are smoothed over time as the compounded volume I_c gets updated. A higher value for the exponential averaging factor ω_0 can improve responsiveness in these cases at the cost of lower denoising. Correct handling of this case will require computationally expensive deformable registration, however possible future extensions could alternatively detect larger deformations in the projection images and locally adapt ω_0 accordingly to compensate.

The processing pipeline we have described in this chapter is perfectly suited to support the real-time requirements of our proposed robotic surgery system. It is has been specifically designed to make as much use as possible of the information contained in the cheaply computable 2D projection maps, which has already been shown to synergize well with the instrument detection and rendering. Furthermore, this information could easily be used to adapt the instrument tracking algorithm described in section 7, completely replacing the detection of candidate points the information from P_{argmax} and M_t.

Considering again the principle requirements we have defined for an integrated visual assistance system defined in section 6, the methodology presented here provides contributions towards the requirements of efficient processing pipelines (**II.**) to provide semantically aware, specialized visualizations (**III.**, **IV.**). The tight integration of processing and rendering provides first insights towards building end-to-end integrated pipelines (**VI.**).

10 Deep Volume Rendering

Contents

In the previous chapters, our work was predominantly focused on developing new methodology that can be applied directly for intraoperative visualization in the retinal microsurgery OR. The visualization concepts presented in Chapters 8 and 9 introduce novel ways to integrate the semantics of the volume content in the visualizations to aid perception of the shown data. In these works, we achieve it by relying on explicit segmentation of structures, in some cases already relying on machine learning to provide those segmentation labels. In developing these visualizations, it became apparent again and again that developing domain-specific volume visualizations still requires a tremendous amount of not only understanding of the domain, but also the technical aspects of how the rendering algorithms work, starting from a detailed understanding of how transfer functions work and are defined, to understanding of illumination models, color perception and a plethora of other aspects.

In this chapter, we present a rather more fundamental line of research on how the DVR pipeline can be improved by integrating learned modules. With the help of these learned components, it becomes possible to invert the workflow of designing visualizations. This means that instead of a forward-oriented design loop of rendering adapting parameters and observing how that change manifests in the rendering until satisfied, users can specify the desired visual outcome and train a rendering model that will provide this outcome.

This chapter presents novel research that has been made available publically as a preprint [217].

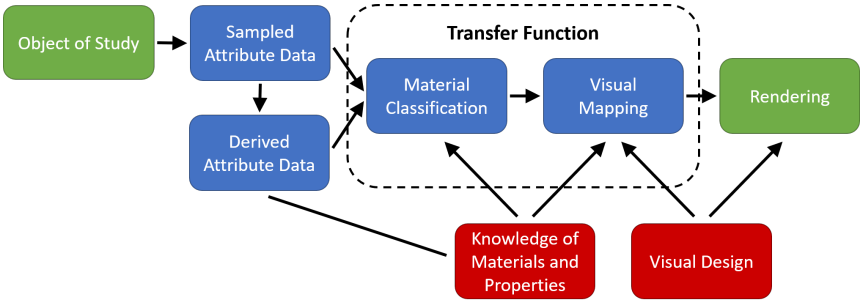


Fig. 10.1. Schematic workflow of classic transfer function design. TF specification requires understanding of the input and derived attributes, internal parameters and the intended visual outcome. Figure reproduced from Ljung et al. [129].

10.1 Introduction

Volume rendering has become an important part of scientific visual computing as an alternative to viewing cross-sections or extracting explicit surfaces from scalar 3D datasets. As a widely researched technique for visualizing structural information in volumetric data sets, many of its aspects have been explored intensively in the last years. It has been used in many application areas involving volumetric data, such as weather simulation, structural imaging in engineering, and medical imaging.

Volume Rendering owes much of its versatility to the use of a transfer function (TF), introduced in the earliest work on DVR [139] as a way of assigning optical properties (color and absorption) to the voxels of a volume. Designing TFs is split into a classification step and a visual mapping step (c.f. Figure 10.1). The classification step derives a semantic based on original and derived voxel attributes and the visual mapping associates visual properties (typically color and opacity). Both steps are guided by domain knowledge and understanding of the visual parameters involved in rendering. Despite years of research on the specific design and parametrization of TFs, obtaining robust TFs that can be applied to several volumes and provide consistent highlighting remains a challenge. This is due in part to the expertise required, as both a technical understanding of the visual parameters and sufficient domain knowledge are needed to design a good TF. The task is complicated further by variations in the data which might cause structurally different areas to hate the same intensities (or general attributes), which complicates the classification step of the TF. To overcome this problem, much of the research in DVR has been dedicated to finding features that can discriminate structures that overlap in intensity space, introducing task-specific solutions that often do not generalize well to other use cases. Furthermore, using additional features introduces the problem of designing a multi-dimensional TF. Specification of 2D TFs already requires a lot of manual interaction and multidimensional TFs specification is a complex problem that in itself motivated many publications [129]. One potential solution to achieve consistent colorization is using explicit semantic labels, provided by semantic segmentation or manual labels. However, training semantic segmentation models requires the annotation of typically hundreds of slices *per volume* to create robust data sets. In medical applications, this has a high cost associated as labeling has to be performed by clinical experts. Furthermore, creating

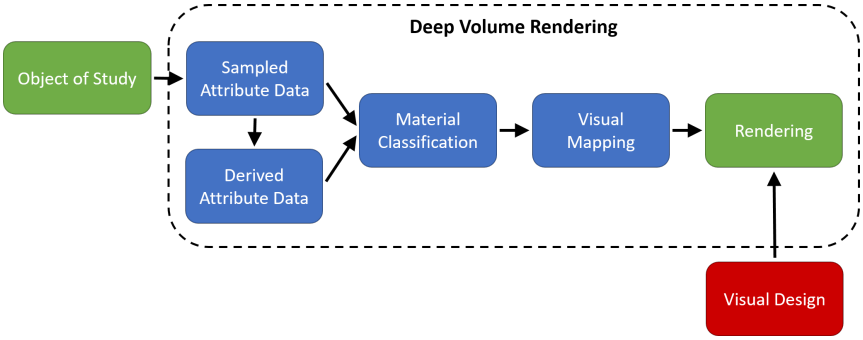


Fig. 10.2. Modified workflow with deep direct volume rendering. Interaction is in image space and only requires specification of the desired visual outcome with less knowledge of the input attributes and materials.

expressive visualizations from the raw data and labels is still not a trivial task and often requires much iteration between domain experts and visualization specialists.

The techniques we present here eliminate the tedious process of finding features and specifying the associated visual properties by replacing the respective steps with learned models. We revisit the classic DVR pipeline and systematically identify the steps that can be supported with learned deep mapping functions. The differentiable nature of DVR allows us to train these models from 2D images, effectively reprojecting 2D annotations into 3D where meaningful features and visual mappings are derived implicitly during training. Our proposed process avoids the time-intensive highlighting of *where* interesting structures are in 3D and instead allows experts to work directly in the rendered image space to indicate *what* they want to see and *how* it should appear. This requires fewer annotations and these annotations can directly reflect the intended visual outcome. Besides direct manual annotation, a wide range of other techniques could also be used to create the training images for our methods. For example, surface meshes could be reconstructed for a limited data set and surface shaders could be used to create a highly customized illustrative rendering. Our end-to-end training can incorporate the illustrative rendering aspects and apply it to unseen input data without the need for explicit surface extraction. Our work is the first to describe neural rendering for scientific volumes while directly modeling the feature extraction and TF mapping steps within the model. Even though our architectures are currently constrained by relatively long training and inference times, clear benefits of our work can already be seen for the medical context.

In the following, we introduce a unified framework for direct volume rendering that incorporates learning of implicit functional mappings by explicitly modeling the DVR process as part of the deep architectures.

Key Contributions

- A generalized formulation of DVR that allows the integration of deep neural networks into parts of the classic pipeline, thus enabling *Deep Direct Volume Rendering* (DeepDVR)
- A set of deep volume rendering architectures which are derived from this formulation by introducing deep networks in several parts of the rendering
- An effective training strategy for models with explicit ray casting we call *stepsize annealing*, validated by experiments
- Experiments to demonstrate of the effectiveness of image-based training of our deep rendering architectures on the tasks of (1) image-based TF specification and (2) learning generalized visual features in volume and image space

The chapter is structured into the following sections: Section 10.2 provides an overview of the related literature. Section 10.3 introduces the mathematical foundation of our *deep direct volume rendering* and presents several DeepDVR architecture we have considered. We also present DVRNET, a novel architecture for multiscale feature extraction in volume rendering tasks. In our experiments (section 10.4), we first perform a detailed analysis of DeepDVR-specific metaparameters and parametrizations for intensity transfer functions. These experiments demonstrate the effectiveness of our *stepsize annealing* training method. In section 10.3, we compare our deep architectures on two tasks: (1) learning to render from manual annotations in image space and (2) learning generalized volume rendering from a multi-volume data set. We discuss the implications of our experiments in section 10.5 and provide concluding remarks in section 10.6.

10.2 Related Work

Volume Rendering and Transfer Functions

Drebin et al. [44] have introduced the first formulation of direct volume rendering, however the GPU-accelerated pipeline described by Krüger and Westermann [114] have played a crucial role defining modern hardware-accelerated ray casting. Multidimensional TFs, leveraging derived attributes to improve classification, have been proposed in variations. Kniss et al. [109] have proposed the use of gradient magnitude and a directional gradient attribute in addition to intensity. Different derived attributes based on for example curvature [106], size [35] or a local occlusion spectrum [36] were also introduced. Ljung et al. [129] provide an excellent overview of recent techniques for TF design in their state of the art report.

The complexity of two-dimensional TF specification has been often dealt with by specifying 2D geometries on the 2D feature histogram [109]. Rezk-Salama et al. [168] have proposed the use of semantic models represented as 2D primitives in the intensity-gradient magnitude

space. In their work, experts can create a basic semantic model which can then be easily adapted by non-experts to a specific data set. Kohonen maps can also be used to reduce the high-dimensional feature space to two dimensions in which the same user interfaces can be used [40]. Schulte zu Berge et al. [184] introduced a simplified approach to specify the importance of multidimensional features by using a predicate weighting for each feature.

Machine learning has already been applied to replace the TF to some extent: Soundararajan et al. [194] use machine learning models to learn a TF from scribbles in the volume domain. They compare five machine learning approaches including a single layer perceptron (SLP), finding random forests to be favorable due to their superior speed and robustness. Approaches have been proposed for modifying TFs in image space via strokes to indicate areas that should be changed in the output image. Ropinski et al. [175] use feature histograms along the rays covered by the strokes to adapt the TF and Guo et al. [81] extended these ideas with novel interaction metaphors for contrast, color and visibility control directly in image space. Ruiz et al. [179] iteratively adapt TFs by optimizing the visibility distribution of local volume features towards a desired target distribution in the output image.

Deep learning has also been used in the context of volume rendering to synthesize novel views or views with different parameters [85, 92], for super-resolution of volume isosurface renderings [220], for compressed rendering of time-varying data sets [97] and for prediction of ambient occlusion volumes [58]. Recent, yet unpublished parallel work by S. Weiss[1] and Westermann[221] demonstrates that the differentiability of DVR can be leveraged to optimize the involved parameters (viewpoint, transfer function and voxel densities) based on image-based loss functions. They show how gradient computation with respect to rendering parameters can be computed efficiently with reduced time and memory consumption. Their work is complementary to the work presented here and offers improvements over the naïve implementation of differentiable DVR our methods are based on.

Neural Rendering

Differentiable rendering has been a subject of recent interest in computer graphics as a building block for machine learning pipelines [104]. Differentiable mesh rasterizers [123, 127, 132, 158] provide interesting solutions to inverse rendering tasks, however they are generally not extendable to volumetric data. Neural rendering [204] is a relatively novel field in which a rendering step is incorporated into the network architecture. This allows for the explicit or implicit incorporation of scene parameters into the training and enables applications like scene relighting, novel view synthesis, facial and body reenactment and photorealistic avatars, among others.

The specific combination of volume rendering with machine learning has been addressed in recent literature: Nguyen-Phuoc et al [154] introduced RenderNet, a deep convnet that performs differentiable rendering of voxelized 3D shapes. Their proposed network consists of a 3D and a 2D convolutional part connected by a novel *projection unit* which combines the features along the viewing ray with an MLP. This enables the network to learn different rendering styles and was also shown to be effective to synthesize novel viewpoints for faces captured from a single viewpoint. Rematas et al. [167] present a controllable neural voxel

[1] Sebastian Weiss is not related to the author of this Dissertation

renderer based on this *projection unit* which produces detailed appearance of the input, handling high frequency and complex textures. More pertinent to scientific volume rendering, Berger et al.[16] have formulated the volume rendering task as an image generation from a given camera and TF. They use generative adversarial networks (GANs) to effectively train a network to memorize a specific volume dataset, representing a differentiable renderer for this volume that is conditioned only on viewpoint and TF. This differentiable renderer can then be used to further explore the latent space of TFs and provide a sensitivity map visualizing areas in the output image affected by specific parts of the TF.

Differentiable rendering based on ray casting has recently gained attention in the context of inverse rendering for scene reconstruction, where different scene representations have been explored: [130] uses *Neural Volumes* to optimize volumetric scene representations from sets of camera images. They use an end-to-end approach to reconstruct an explicit color+opacity volume that is then rendered from the known viewpoints and compared to the reference images. Instead of learning an explicit volumetric representation, Scene Representation Networks [188] encode both geometry and appearance into the network itself as a mapping of 3D location to feature vector. The feature vector of each position encodes both the signed distance to a surface, as well as features that are later decoded to the final surface color. [126] propose an efficient differentiable sphere-tracing algorithm to render implicit signed distance functions. [144] train a network that is conditioned on a 3D position and viewing direction and returns an RGBA tuple, thus encoding a *Neural Radiance Field* in the network. This radiance field can be sampled at arbitrary points to create novel photorealistic viewpoints from a set of images while accurately representing complex geometry and materials. Niemeyer et al. [157] follow a similar approach to learn deep implicit representations of shape and texture without 3D supervision. Although many of these works include a volume ray casting step within the training pipeline, neural rendering has not yet been proposed for scientific volume visualization.

10.3 Learning Visual Feature Mappings

The beneficial properties of the widely used classical emission-absorption model [139] for DVR provide a strong potential for inverse rendering tasks in scientific volume rendering. Our considerations are based on the mathematical formulation of ray casting as introduced in section 2.3.

Numerical integration of the ray with a fixed sampling rate can lead to aliasing artifacts. A common approach to reduce this is to use *ray jittering* [38] which applies a random offset from a uniform distribution $t_o \sim \mathcal{U}(0, t_{j,\max})$ in the ray direction $\mathbf{d}$ to each pixel ray as

$$\mathbf{x}(t) = \mathbf{x_0} + (t + t_o)\mathrm{d} \tag{10.27}$$

where the jitter magnitude $t_{j,\max}$ is usually chosen as Δt.

For a more in-depth discussion of the derivations, we refer the interested reader to the course notes by [78].

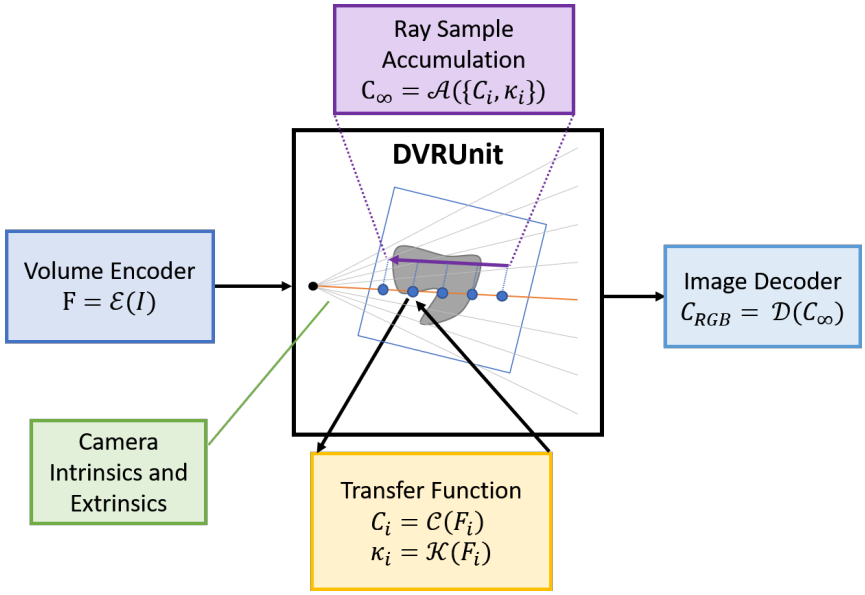


Fig. 10.3. Schematic overview of the proposed DVR Unit based on our generalized DVR formulation. Our architectures derive from this by implementing these functional mappings with deep neural networks.

In order to perform image-based optimization of the rendering algorithm, we require an extensible differentiable mathematical model of volume rendering. We generalize the well-known emission-absorption volume rendering equations to an arbitrary number of color channels. We also introduce replaceable functions for feature extraction, color and opacity TFs, composition and color space decoding, which can be then implemented with learning-based approaches.

10.3.1 Generalized Direct Volume Rendering

Our novel deep volume rendering method is based on the observation that the DVR algorithm is essentially independent of the color space used, and thus can be performed in a latent color space of arbitrary dimensions. The emissive component c is commonly represented as a linear RGB triplet, however observing the volume rendering integral (2.1) and especially the front-to-back blending use to approximate it, 2.4, it is possible to perform the blending in any (linear) color space. We formulate our deep ray casting algorithm using a generic n_F-dimensional input feature space $\boldsymbol{F} \in \mathbb{R}^{n_F}$ and an n_c-dimensional color space $c \in \mathbb{R}^{n_c}$. Our deep volume rendering pipeline consists of three distinct parts: (1) feature extraction, (2) deep volume rendering by volume sampling and subsequent sample accumulation and (3) image decoding (c.f. Figure 10.3).

The purpose of input feature extraction is to derive further features to discriminate between different classes in order to assign different optical properties. We formalize this by introducing a *volume encoder* function $\mathcal{E}(I) : \mathbb{R}^{n_I,H,W,D} \rightarrow \mathbb{R}^{n_F,H,W,D}$ that transforms the scalar input volume $I(\mathbf{x})$ to an n_F-dimensional feature space:

$$F(x) = \mathcal{E}(I(\mathbf{x})) \tag{10.28}$$

We split the TF into two parts and introduce $\mathcal{C}(F) : \mathbb{R}^{n_F} \rightarrow \mathbb{R}^{n_C}$ and $\mathcal{K}(F) : \mathbb{R}^{n_F} \rightarrow \mathbb{R}$ which compute the emission $c \in \mathbb{R}^{n_C}$ and absorption coefficient κ, respectively, from the feature space. This yields the emission C_i and absorption A_i for each sampling point:

$$F_i = F(\mathbf{x}(i\Delta t)), \tag{10.29}$$

$$C_i = \mathcal{C}(F_i)\Delta t, \tag{10.30}$$

$$A_i = 1 - e^{-\mathcal{K}(F_i)}\Delta t. \tag{10.31}$$

The samples (C_i, A_i) of a ray are then combined with a *accumulation function* $\mathcal{A}$ to a final ray color:

$$C'_\infty = \mathcal{A}\left(\{(C_i, A_i)|i\}\right). \tag{10.32}$$

Within this arbitrary n_C-dimensional color space, alpha blending equations can still be evaluated as described per equations 2.4-2.5. However, $\mathcal{A}$ could also represent illustrative or importance-based compositing operations [22, 39], or even be replaced with learned compositing methods in the future.

The resulting image then consists of the n_C-dimensional projected features C_∞ and an alpha channel A_∞. We therefore introduce an *image decoder* function $\mathcal{D}(F)$ to transform the rendered features back into an RGB color space via

$$C_\infty^{RGB} = \mathcal{D}(C'_\infty). \tag{10.33}$$

10.3.2 TFs based on MLPs

In scientific volume visualization, TFs are routinely used to map from the voxel feature domain $F(x)$ to the optical properties (emissive color c and absorption coefficient κ). This mapping strongly influences the visual aspects of the rendering (c.f. Figure 10.1), controlling both which structures are relevant (*classification*) and how these structures appear visually (*visual mapping*). As such, the TF is usually parametrized such that users can modify and tweak it to a specific data set in order to achieve the desired effect. In practice, TFs are often represented as lookup tables or through simple primitives like trapezoid and parabolic functions [168] or a sum of Gaussian functions [110]. Manual specification of traditional TFs is already tedious in one- and two-dimensional feature spaces and has to rely on dimensionality reduction techniques for higher dimensions to make the problem even tractable from a user interaction perspective [40]. Yet, TFs mathematically are merely a mapping from an n_F-dimensional input to a 4-dimensional (RGBκ) output space for the emission+absorption DVR model and can therefore be approximated with multi-layer perceptrons (MLP).

Multi Layer Perceptrons are a type of neural network that use a layered design. Each layer has an N_I-dimensional input vector $\boldsymbol{x}$ and an N_O-dimensional output vector $\boldsymbol{y}$. Each component y_i is computed as the weighted sum of all input values and applying a non-linear activation function $f(x)$:

$$y_i(\mathbf{x}) = f\left(\sum_{j=1}^{N_I} w_{i,j}x_j + b_i\right) \tag{10.34}$$

where $w_{i,j}$ are the learned weights for each input and b_i is a learned bias value. Many different activation functions have been used for deep neural networks[3], however in this work we will only use the classic ReLU and sigmoid activation functions.

Note: Transfer Function Derivatives

To perform gradient descent in a typical deep-learning optimization loop, the derivative of the transfer function with respect to its parameters as well as its input is required. Lookup functions are typically implemented via a 1D interpolated texture lookup, so it is not intuitive how this can be implemented as a differentiable method. Given the k color values of the lookup table as c_k, the transfer function lookup can be expressed in using the *hat function* $\Lambda(x) = \max(1 - 2|x|, 0)$:

$$TF(x) = \sum_k c_k \Lambda(x - k + 0.5).$$

From this, the partial derivatives can be trivially using the partial derivative of the hat function:

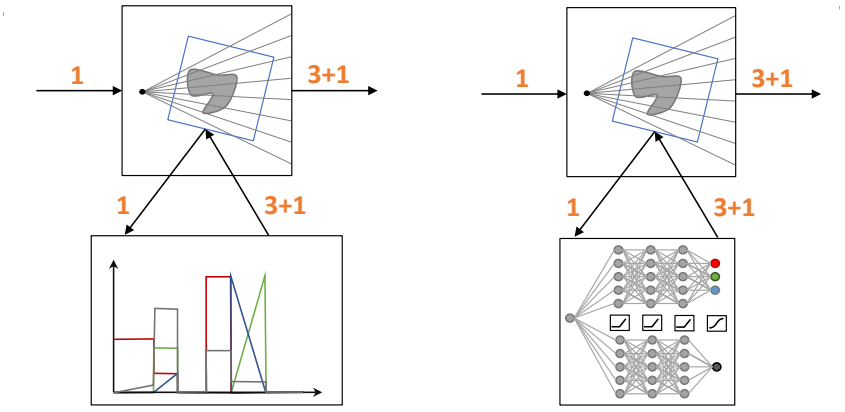
$$\frac{\partial \Lambda(x)}{\partial x} = \begin{cases} -2 & 0 < x < 0.5 \\ 2 & -0.5 < x < 0 \\ 0 & \text{otherwise.} \end{cases}$$

By replacing the traditional TF representation with an MLP (Figure 10.4), we eliminate the need for directly modifying the mapping function manually, as the MLP parameters for the most part do not allow for meaningful direct manual manipulation. Instead, these parameters are optimized mathematically given example data. Given the differentiability of the DVR in the formulation outlined above, these parameters can be optimized as part of the proposed end-to-end training procedure. For simple 1-dimensional input feature spaces this is not very important. However, lookup tables quickly grow impractical for higher-dimensional feature spaces while MLPs can scale more gracefully to higher input dimensionality.

10.3.3 Deep Direct Volume Rendering

The generalized reformulation of the direct volume rendering algorithm above was also motivated by the goal to eliminate the explicit design of application-specific features in a general way by using a convolutional neural network (CNN) for the feature extractor $\mathcal{E}$. Our algorithm also allows for creating architectures that perform ray casting in a latent higher-dimensional color space and then transform this high-dimensional image to the RGB color space.

In the following, we outline a set of architectures derived from our general formulation of direct volume rendering in Section 10.3.1. We have designed several architectures based on replacing the generalized functions for feature extraction $\mathcal{E}(I)$, color and opacity TFs $\mathcal{C}(F), \mathcal{K}(F)$ and the image decoder $\mathcal{D}(C)$.



(a) Lookup TF (b) MLP-based TF

Fig. 10.4. Transfer function representation alternatives. **(a)** Lookup TFs are based on keypoint definitions and typically resampled into a 1D texture. **(b)** MLPs can be used as an alternative to represent the same function. *Orange numbers indicate the number of channels*.

Our architectures, summarized in Table 10.1 and Figure 10.6, are progressing in complexity by replacing more of the functions with deep neural networks. For comparison, we also include RenderNet [154] as a baseline in our experiments as, among previously published works, this architecture is most closely related to our concepts. RenderNet resamples the input volume to view space using the perspective camera transform, such that the X-axis of the resampled volume corresponds to the ray directions and the YZ-axes correspond to the image coordinates. This resampled volume is then processed by a 3D convolutional network (CNN) to extract semantic features, reducing the spatial dimension in the process. This step corresponds to our *input encoder* and *opacity and color TFs*. RenderNet uses an MLP to project all voxels along the X-axis to a single high-dimensional feature vector. This *projection layer* is analogous to our generalized *accumulation function* $\mathcal{A}$. The resulting 512-channel 2D projection is then upsampled to the target resolution using a 2D CNN, similar to our *image decoder* function $\mathcal{D}(C)$. There are some key differences between RenderNet and our approach of deep direct volume rendering: Firstly, the view space resampling in RenderNet is performed at the beginning of the pipeline whereas in our approach, the volume encoder extracts features in object space, making them independent of the camera parameters. Secondly, our formulation uses a more explicit modeling of occlusion through the absorption coefficient and alpha blending. Thirdly, we avoid excessive downsampling in the spatial dimensions by using specific architectures that employ skip connections.

We introduce four novel architectures based on these considerations: three VNet-based architectures (VNET-4-4, VNETL-16-4, VNETL-16-17) and DVRNET, a novel multiscale rendering architecture.

VNET4-4 The first architecture was created by implementing the *volume encoder* $\mathcal{E}(I)$ with a 3D convolutional deep neural network. We chose to use an existing, well established multiscale encoder-decoder architecture called VNet[146] which has been shown to perform well for binary volumetric image segmentation. We adapt this network to a four channel output to

	$\mathcal{E}(I)$	$\mathcal{C}(F)$	$\mathcal{K}(F)$	$\mathcal{A}(\{(C_i, A_i\})$	$\mathcal{D}(C)$	Parameters
Lookup TF	$\mathcal{I}_1$	Lookup	Lookup	Alpha	$\mathcal{I}_3$	~1 K
RenderNet	~~~~~~~3D CNN ~~~~~~~			MLP_{512}	2D CNN	~226 M
VNET4-4	VNet_4	$\mathcal{I}_3$	$\mathcal{I}_1$	Alpha	$\mathcal{I}_3$	~45.6 M
VNETL16-4	VNet_{16}	MLP_3	MLP_1	Alpha	$\mathcal{I}_3$	~12.3 M
VNETL16-17	VNet_{16}	$\mathcal{I}_{16}$	MLP_1	Alpha	MLP_3	~12.3 M
DVRNET*	VNet Blocks	$\mathcal{I}$	MLP	Alpha	UNet Blocks	~25.0 M

Tab. 10.1. Summary of novel architectures and comparison to baseline methods. $\mathcal{I}_N(x) = x$ is an N-dimensional identity function. "Alpha" designates alpha blending. * DVRNet uses DVR at multiple scales and does not fit directly into this categorization.

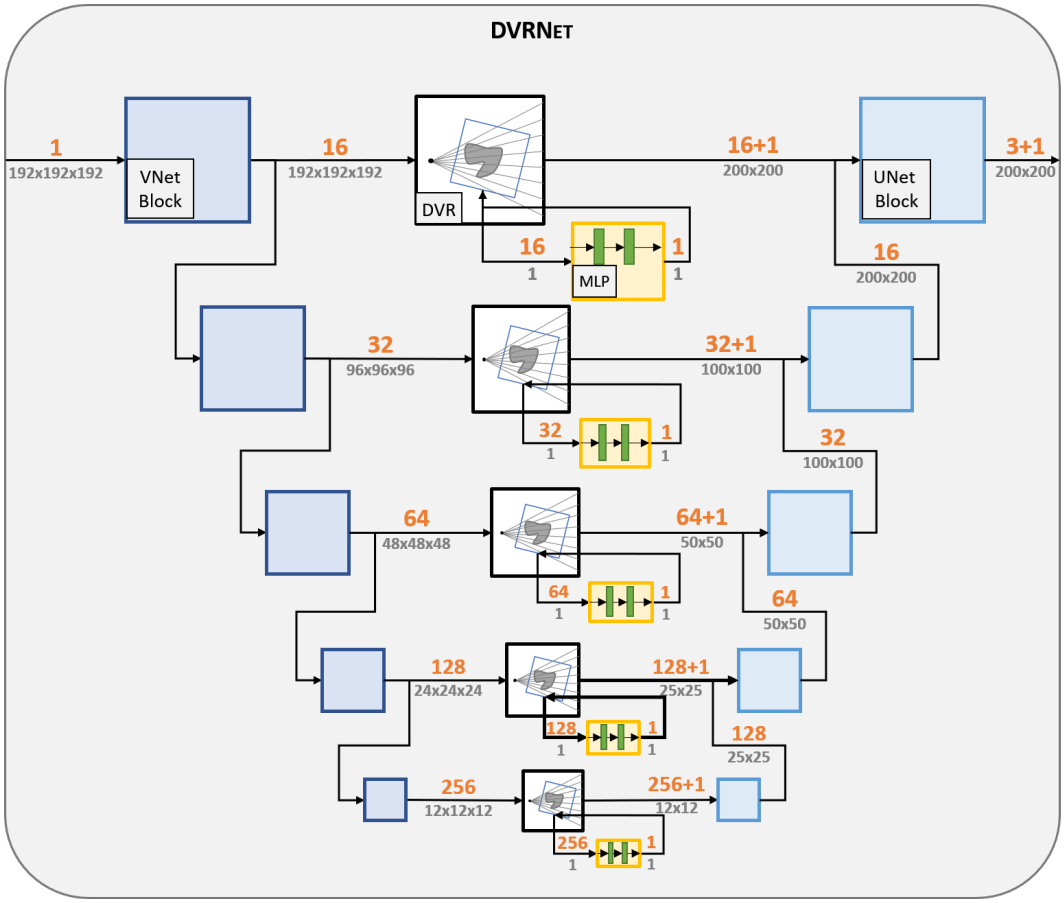


Fig. 10.5. Schematic overview of our novel DVRNET architecture. Inspired by UNet and VNet architectures, DVRNET features a multiscale volumetric encoding part and a multiscale 2D decoder. Corresponding encoder-decoder levels are connected with a DeepDVR module, replacing the skip connections of the classic UNet/VNet architectures. Indicated dimensions are for the training data set, the convolutional and DeepDVR layers support arbitrary input and output dimensions during and after training. "*+1*" *indicates a separately handled opacity/alpha channel.*

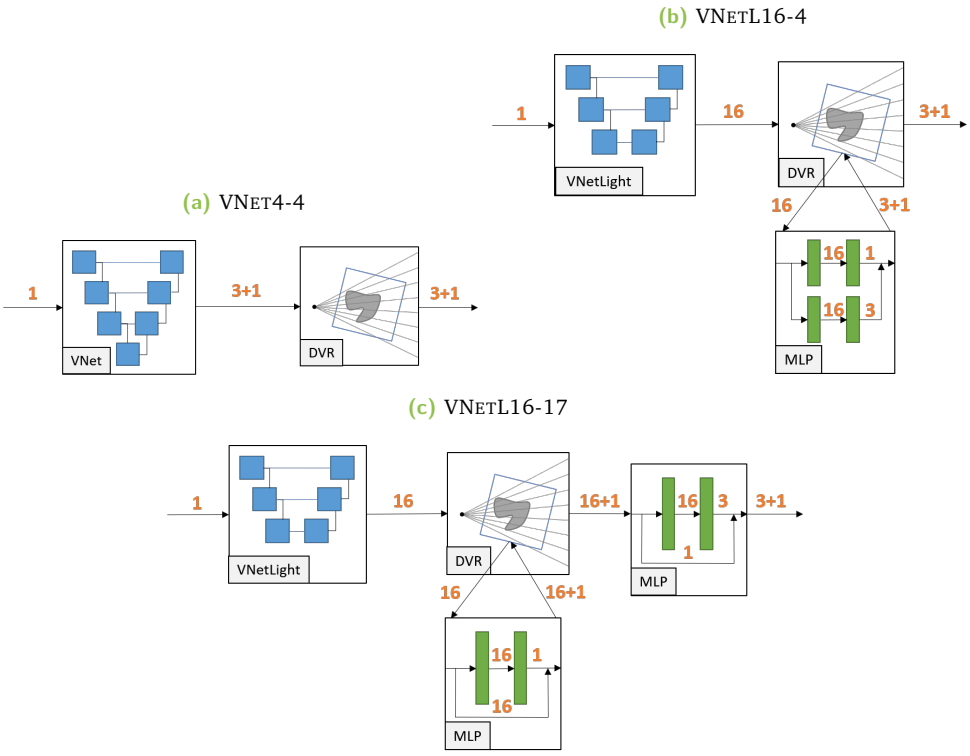


Fig. 10.6. Our three VNet-based model architectures with increasing complexity. VNet4-4 uses a deep encoder network to directly infer the RGBκ volume to be rendered. VNetL16-4 creates 16 semantic channels in the encoder which are mapped to color and opacity via an MLP-based TF. VNetL16-17 instead uses an MLP to map these channels to 16 color space channels, which are then decoded with a final MLP on the 16-channel output image. "*+1*" *indicates a separately handled opacity/alpha channel.*

obtain an RGBκ volume directly. This essentially implements a *pre-classified* volume rendering, for which [59] have discussed further rendering optimizations we do not apply here, but which could be used outside of training.

VNetL16-4 This architecture is extended from the previous by replacing the VNet network with a simplified "VNetLight" variant with one less level of downsampling to reduce the number of parameters, and by producing 16 semantic channels instead. Additionally, we add MLP-based trainable color and opacity TFs. For each of the two TFs, an MLP is defined that feeds the 16-channel input to a 16-channel hidden-layer with a ReLU activation and then produces either 3 channels (RGB) or a single channel (κ) using a sigmoid activation function. Again, no *image decoder* is required as the image is produced in the RGB color space.

VNetL16-17 This architecture performs compositing in an abstract 16-channel color space. As before, a VNetLight architecture implements the *volume encoder* to produce a 16-channel volume. For $\mathcal{K}(F)$, the same MLP architecture as with VNetL16-4 is used. Instead of a color TF, the 16 semantic channels are directly used in compositing (i.e. $\mathcal{C}(F) = F$). To decode the resulting 16-channel image (with one additional alpha channel), we use another MLP with one ReLU-activated 16-channel hidden layer and a sigmoid function in the output layer, yielding the 3-channel RGB image whereas the opacity is used directly as produced from the DVRUnit.

DVRNet Our architecture is inspired by the encoder-decoder architectures that have shown great success in image segmentation due to their ability to use both local and global semantics by their multiscale nature. Our architecture, visualized in Fig. 10.5, is based on the idea that the volumetric multiscale encoder part of a VNet can be connected with the upsampling decoder part of a U-Net by inserting a DVRUnit into the skip connections to perform a projection from 3D to 2D feature space. Each level viewed on its own is a volume ray casting network similar to VNetL16-17, rendering a potentially spatially downsampled volume in a latent color space at a reduced image resolution. These lower-resolution multichannel images are then combined in the U-Net decoder blocks to produce the final 2D image at the desired resolution.

This architecture was designed to achieve a similar multiscale classification power as using VNet directly as an encoder, without the need to perform full volumetric upsampling, performing the upsampling step in image space instead.

10.3.4 Stepsize Annealing

The sampling rate s has a major influence on the computational footprint of the algorithm: it linearly scales the number of samples classified and composited for all rays. When optimizing weights with automatic differentiation, this scaling also applies to the backpropagation step and memory consumption. Given that increasing the sampling rate severely increases training time at diminishing returns, we propose a novel training strategy for a more efficient training of differentiable volume rendering models: In order to benefit from the better convergence efficiency at low sampling rates, we propose a strategy that varies the sampling rate in a range $[s_l, s_h]$ across the epochs during training relative to the training progression where e is the current epoch and E the total number of epochs:

$$s(e) = s_l(1 - (e/E)^2) + (e/E)^2 s_h. \tag{10.35}$$

This progressive supersampling of the volume during training results in a fast convergence with low sampling rates in first epochs for a rough approximation and increases the sampling rate towards the end to refine the optimized TF.

10.4 Experiments

Our differentiable volume rendering enables the end-to-end optimization of rendering parameters, yielding models that do not require tedious manual manipulation of transfer function parameters.

In the following sections, we evaluate the effectiveness of the presented concepts. We first present experiments regarding the direct learning of the functional mapping $F \to (C, \kappa)$ in image space, comparing the performance of TF representations. In these experiments, we also evaluate how the hyper-parameters specific to volume rendering affect training and testing performance. In the second set of experiments, we compare the our deep architecture variants

for volume rendering by training on manually adapted reference images for a single volume (section 10.4.2) and generalizing across multiple volumes to create a renderer that is robust against inter-volume variations (section 10.4.3).

10.4.1 Image-Based TF Optimization

In this section, we evaluate the benefits of optimizing 1D lookup TFs and MLP-based TFs in a task of image-based TF definition.

As a straightforward benchmark for our methods, we reconstruct the TF from a set of images for a single volume such that the images produced by an optimized TF are as close as possible to the reference images. While a task like this does not directly have practical applications as-is, it serves here as an evaluation framework with well-defined, perfect ground truth to investigate the training behavior of image-based TF optimization with respect to several parameters.

Experimental setup

We train two different TF representations: a classic lookup TF where the 256 individual elements are optimized, and an MLP representation with two hidden layers. Both representations have a similar number of trainable parameters and we have confirmed that the MLP is deep enough to represent the TFs used in this experiment. The goal of this experiment is to analyze the influence of hyperparameters that are specific to DeepDVR in order to optimize the training process. We evaluate how training with different sampling rates affects training performance and validate our novel *stepsize annealing* scheme. For the Lookup TF, we further analyze on whether ray jitter (c.f. Equation 10.27) has an effect on training. In summary, our experiment conditions are: *model* (Lookup, MLP), *sampling rate* (fixed $s \in \{0.25, 0.5, 1.0, 2.0, 3.0\}$ vs. stepsize annealing with $s_l = 0.1, s_h = 2.0$) and *ray jitter* (yes/no).

Data sets We create five training data sets from five different volumes sourced from the volume library[177] (c.f. Figure 10.7, top row) for each of which we manually designed a 1D TF. Each of the volumes was converted to floating point, resampled and padded to an isotropic voxel resolution of $256 \times 256 \times 256$, and then a training set of 25 images was created with a normal DVR renderer using the manually defined TF, with an additional 7 views as a validation set.

Metrics We use structural similarity (SSIM)[213] to assess image quality after every epoch. To evaluate the perceptual quality of the produced images, we report two perceptual similarity metrics: The *Fréchet Inception Distance* (FID) score [90] and *Local patch-wise image similarity* (LPIPS)[236]. Both metrics are based on auxiliary, pre-trained neural networks to evaluate perceptual similarity instead of per-pixel comparisons. The specified sampling rate s in the charts is exclusively used for training. During testing, we render all images using $s = 3.0$ to provide a fair comparison with the ground truth. We also measure the total time required for training all epochs on our hardware as an indicator for the computational scaling.

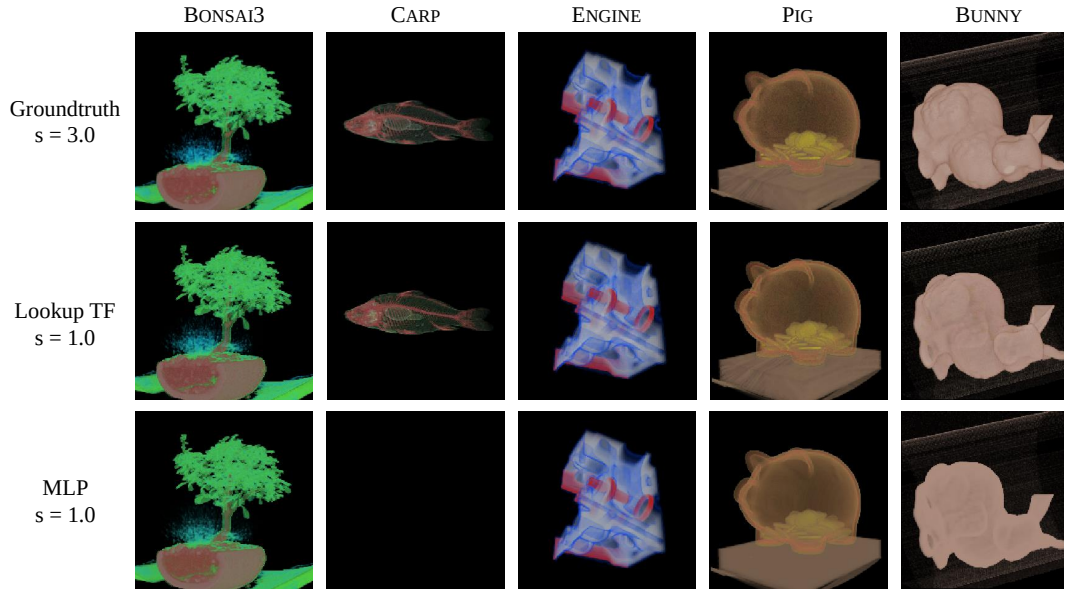


Fig. 10.7. The five volume data sets rendered with manually adapted reference TF (top row), reconstructed lookup TF (middle row) and deep TF represented as an MLP. s: sampling rate.

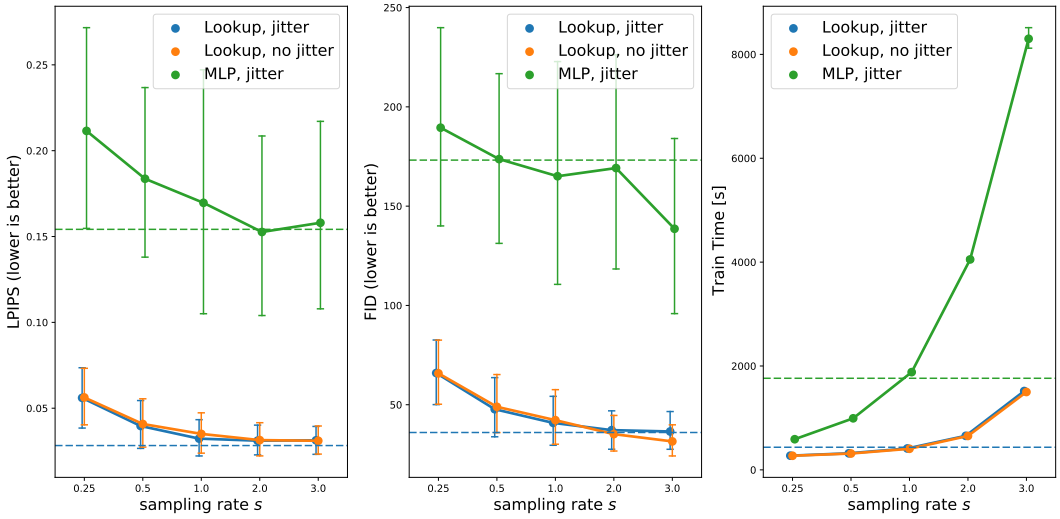


Fig. 10.8. Training behavior with respect to sampling rate. Graphs show average metrics and 95% CI, horizontal lines indicate results for our proposed *stepsize annealing* training strategy. Each data point summarizes 25 trainings on the 5 data sets.

Training We use the Adam [107] optimizer ($\beta_1 = 0.9, \beta_2 = 0.99$) at a learning rate of 0.3 for 100 epochs. SSIM is tracked for every epoch and we retain the model with the highest value. We performed MLP training with the same set of training conditions, however only train with ray jittering enabled. We use a fixed learning rate of 0.3 for Lookup and 0.05 for MLP models in all runs. The batch size was chosen manually for each of the sampling rates as higher sampling rates require more GPU memory. We used batch sizes of $\{12, 6, 3, 2, 1\}$ for the sampling rates $\{0.25, 0.5, 1.0, 2.0, 3.0\}$ respectively. Our models are implemented with with Pytorch. All trainings were performed on an NVidia Tesla V100 32GB, 2x Intel Xeon Gold 5120 (2x14 cores) and 256 GB of system RAM.

We refer the reader to our supplementary document, where we provide more details on the training of these experiments as well as more detailed results.

Results

For each of the five datasets, we repeat training five times for each condition with random initialization and report the average across these 25 runs in Figure 10.8. The best of five runs at training $s = 1.0$ can be observed in Figure 10.7.

From Figure 10.8, it can be seen that training time increases with the sampling rate, however the perceptual similarity measures show diminishing returns when increasing the sampling rate beyond 1.0. This is the main motivating argument for our *stepsize annealing* strategy, which is shown as horizontal lines in the plots. For both models, *stepsize annealing* produces visual results comparable to training with a fixed sampling rate of $s = 2.0$ however at an average reduction of 33.4% in training time compared to a constant sampling rate of $s = 2.0$.

Comparing the results with and without jittering there is a slight improvement for low sampling rates (0.25, 0.5) when using jittering. However, the benefit of ray jitter vanishes with higher sampling rates. We assume that ray jittering improves the results for low sampling rates because jittering causes the same ray to cover different locations in 3D space which are otherwise skipped over at low sampling rates. Thus, using ray jittering when undersampling the volume increases the range of intensity samples taken into account across the epochs.

From our results it is evident that the MLP parametrization performs considerably worse than the trained Lookup table. Average training time is increased by a factor of 2.0 (for $s = 0.25$) to 6.19 (for $s = 2.0$) due to the more expensive evaluation of the MLP layers. The visual metrics show a very high variability which we attribute to the random initialization of the MLP weights, showing a high dependence on the initialization.

The qualitative results in Figure 10.7 (bottom row) show that in principle the MLP produces similar rendering for most data sets. Especially the BONSAI3 and ENGINE data sets perform well. PIG and BUNNY exhibit slightly stronger differences and CARP is a clear failure case.

The CARP data set results highlight a general limitation of image-based TF optimization: once the TF reaches a fully transparent state, it cannot recover from it. Closer inspection of the intermediate results during training revealed that the MLP converges to a black image after 5-6 epochs, with a constant loss in the remaining epochs. Due to the use of MSE as a loss function and the very dark and highly transparent target images, a black image caused by

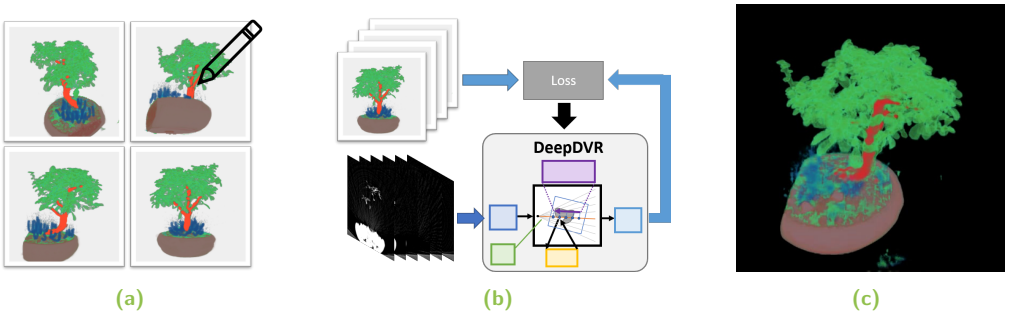


Fig. 10.9. As an example application, we train deep neural networks that explicitly model the feature extraction, classification/visual mapping and compositing of classic DVR for scientific volume rendering. a Our models are trained end-to-end from images that can, for example, be created by domain experts who can now directly specify the desired appearance in view space. b Our DeepDVR architectures model a generalized DVR pipeline and can efficiently learn the visual features required to reproduce these images. c The trained models easily generalize to novel viewpoints, can process volumes of arbitrary dimensions and create high resolution rendered images.

a fully transparent TF is a strong local minimum of the loss function. However, once the resulting images are fully transparent, gradient back-propagation is unable to recover to a non-transparent image: because no samples along a ray are contributing to the final color, no gradient update will be propagated through the samples into the TF and thus the optimization will be stuck in the local minimum. Lookup TFs in general are susceptible to the same problem but there is less incentive to subdue all opacity parameters κ_i to zero at the same time due to the more targeted way parameters affect the intensity mapping in this parametrization.

Overall, our results suggest that replacing TFs with equivalent MLPs is not effective for scalar-valued feature spaces. Nonetheless, MLP-based TFs can still be feasible for higher-dimensional feature spaces where both manual specification and storage of lookup-table-based TFs become impractical.

10.4.2 Learning from User-Adapted Reference Images

To demonstrate the power of deeper architectures to learn complex classification features, we define a task that involves learning from user-adapted images in a workflow as shown in Figure 10.9. This is similar to the scribble-based TF interactions that were proposed in previous work [81, 175], where users can modify the TF via interactions in the 2D image space.

We have created two new datasets based on the original BONSAI3 and PIG data sets consisting of 32 images of 200×200 resolution. The 32 reference images were rendered with an initial intensity-based TF and subsequently edited in an image editor to create an exemplary semantic colorization: For creating the BONSAI-H data set, the spurious green structures from the CT scanning table were removed, the color of the flower pot and the tree's stem were adapted to consistent brown and orange tones and the small grass at the base of the tree was colored blue. In the PIG-H data set, the visibility and brightness of the coins was adapted to a bright yellow and the coin slot at the top of the piggy bank was colored green. Manually designing a

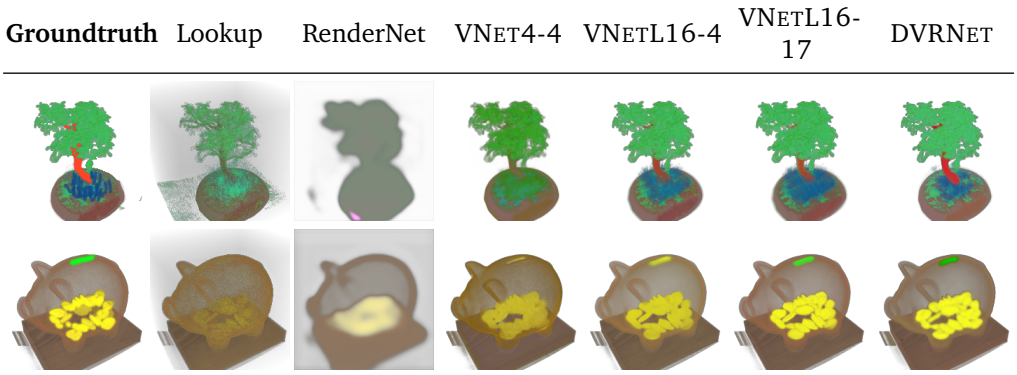


Fig. 10.10. Results of our deep models trained on our manually adapted training images BONSAI3-H (top) and PIG-H (bottom). Results show a single view of the validation set, more examples can be found in the supplementary document. Images best viewed in color at high resolution.

traditional (multidimensional) TF to achieve these visual effects would be extremely difficult to achieve even with well-designed application-specific input features.

Experiment setup

We optimize a rendering model for one specific volume and as such, the resulting rendering model does not necessarily generalize to other data sets in this case.

In this task, we used a combination of MSE and SSIM as our loss function as we found the SSIM in the loss function helpful to recreate finer details:

$$\mathcal{L}(Y, Y_{\text{pred}}) = \mathcal{L}_{\text{MSE}}(Y, Y_{\text{pred}}) + (1 - \text{SSIM}(Y, Y_{\text{pred}})). \tag{10.36}$$

We use the same data split as in Sec. 10.4.1 and train for 200 epochs. We adapted learning rate and batch size for the models in preliminary experiments, maximizing the batch size with respect to the available 32GB of GPU memory for each model. Lookup (bs=32, lr=0.03) and RenderNet ($bs = 10$, $lr = 0.05$) can train with higher batch sizes and learning rates while all VNet* variants and DVRNET are restricted to $bs = 2$ and a $lr = 0.003$.

Results

The qualitative results in Fig. 10.10 clearly show that optimizing the lookup table representation cannot find an adequate TF, instead converging to minimum which achieves only the most general desired visual qualities. The RenderNet architecture evidently is also not suitable for this task. The excessive overblurring in image space is likely caused by the low image resolution in which the projection is performed, where spatial structures in the image can only roughly be reconstructed from the feature vector. Also, the lack of explicit opacity handling requires the model to learn a similar operation in the MLP projection layer, a task that is seemingly too complex to learn even for the with the higher number of learnable parameters of this model. This leads to inconsistent coloring between different views. VNET4-4 manages to assign the general semantics of the green leaves, the flower pot and the yellow coins. It however missed the different coloring of the tree stem and the grass, and cannot reproduce the brightness of the coins inside the piggy bank. VNETL16-4 and VNETL16-17 produce very

Tab. 10.2. Results training our deep architectures on the manually adapted data sets BONSAI3-H and PIG-H. [1]SSIM was used as part of the loss function.

	BONSAI3-H			
	LPIPS ↓	FID ↓	SSIM ↑[1]	Time
Lookup	0.2919	234.47	0.530	6m
RenderNet	0.4943	274.48	0.268	1h 33m
VNET4-4	0.1611	208.04	0.828	8h 32m
VNETL16-4	0.1001	148.80	**0.923**	9h 45m
VNETL16-17	0.1031	170.67	0.921	9h 46m
DVRNET	**0.0800**	**136.52**	0.913	2h 25m
	PIG-H			
	LPIPS ↓	FID ↓	SSIM ↑[1]	Time
Lookup	0.2481	163.74	0.637	6m
RenderNet	0.3368	274.04	0.430	1h 33m
VNET4-4	0.1940	156.95	0.872	8h 8m
VNETL16-4	0.1247	112.20	0.928	9h 54m
VNETL16-17	**0.0830**	106.51	**0.932**	9h 53m
DVRNET	0.0990	**93.36**	0.927	2h 36m

similar results on the BONSAI3-H data set and only differ on the PIG-H results. Both models successfully manage to emphasize the region of the coin slot, however only VNETL16-17 also correctly assigns green color to it. DVRNET performs well on both data sets and reproduces the intended visual attributes in both cases. It picks up better on the distinct coloring of the tree stem, yet has problems reproducing a fully consistent green coloring from all views in the case of the coin slot.

The rough ordering from the qualitative results is reflected in the quantitative results summarized in Table 10.2. DVRNET and VNETL16-17 perform very similar regarding LPIPS and FID metrics whereas VNETL16-4 performs best in terms of SSIM on the BONSAI3-H data set.

In terms of training time, DVRNET shows considerably better speed compared to the full VNet* variants, which can be attributed to the reduced cost of performing the upsampling in image space instead of in volume space. The results of this experiment show that the incremental additions to the models represented by the increasing complexity of VNET4-4, VNETL16-4, VNETL16-17 and DVRNET, successfully improve the perceptual capabilities of the rendering models. Overall, DVRNET seems to perform best in this task when also considering the training performance.

10.4.3 Generalized Rendering Models

In the previous experiment, we have established that the architectures are, in principle, capable of learning meaningful rendering features purely from data provided in image space. In this section, we perform additional experiments showing the benefits of the deep architectures we have designed. To this end, we use a clinical data set with high inter-patient variations, which is a common scenario yet hard to define robust TFs for.

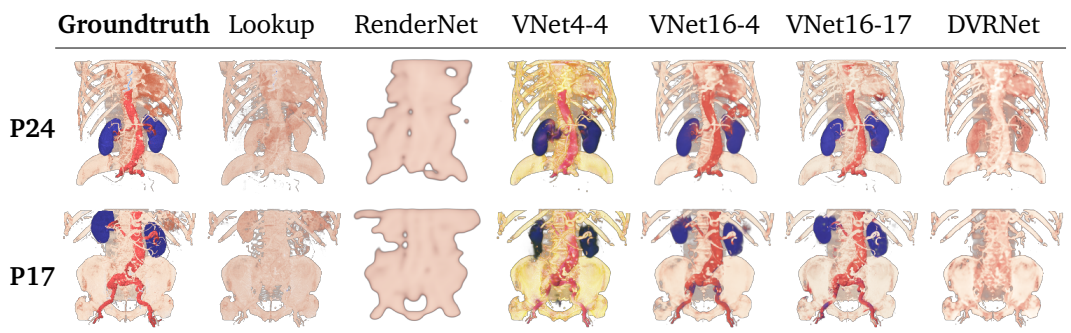


Fig. 10.11. Results of our deep models trained on the KIDNEY data set. Results show the frontal view for two of the six patients from the test set. Images are best appreciated in color and at high resolution.

Experiment setup

The training objective is to optimize rendering models that work consistently between subjects. The volume data consists of abdominal angiographic CT volumes (CTA) from 27 subjects cropped and resampled to an isotropic resolution 192×192×192, together with an equally sized label volume indicating the main abdominal artery and the kidneys with separate labels. The 3D labels were exclusively used to generate the 2D training data and not at as an input to the models, as labels cannot be assumed to exist in a normal clinical setting for a new patient. For each of the volumes, an intensity-based TF preset was adapted manually by shifting the intensity range to produce consistent images showing the bones and contrasted vessels.

From this volumetric data, we create two separate data sets via a shader-based DVR implementation: KIDNEY uses the labels to emphasize kidneys and the labeled artery by applying distinct colors and boosting the opacity. In SHADED, we do not use the label volume but instead use an ambient and diffuse shading model with secondary shadow rays as discussed by [175] to illuminate the ground truth images. In these images, the light source is always at the top-left of the camera (view dependent illumination), a common illumination setup for medical DVR. For both KIDNEY and SHADED datasets, training images are generated at a resolution of 200×200 for a fixed set of 16 viewing directions on a sphere around each volume. The data set was split randomly by patient, using 15 patients for training and 6 patients each for validation and testing. Our two data sets therefore consist of 240 images in the training set and 96 images in the validation and test set, respectively.

We used the same $\mathrm{SSIM} + \mathrm{MSE}$ loss function as in the previous experiment. We have trained all networks on both data sets using the Adam [107] optimizer ($\beta_1 = 0.9, \beta_2 = 0.99$) and have used the same learning rate and batch size combinations as in section 10.4.2. All models except RenderNet were trained with *stepsize annealing* with $s_l = 0.1, s_h = 1.0$.

Results

The results are summarized in Table 10.3, reporting LPIPS, FID and SSIM. However, SSIM was used in the optimization and should therefore be interpreted accordingly. A qualitative comparison for two patients in the test set can be seen in Figure 10.11. Results on all six patients in the test set can be found in the supplementary document.

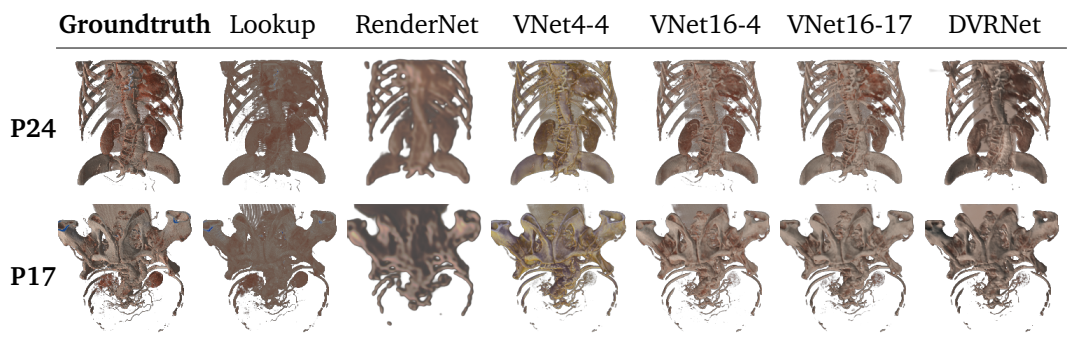


Fig. 10.12. Training results for the SHADED data set. Results show two different viewing directions for two patients from the test set to demonstrate how the different models are able to learn view-dependent illumination. Images are best appreciated in color and at high resolution.

Inspection of the resulting images reveals that the optimized Lookup TF that was included as a baseline here misses the primary objective in both data sets: neither can it colorize the structures correctly nor is it able to represent any shading information. Both of these effects are expected since it is impossible to discriminate bone, vessel and kidney structures purely based on intensity and shading needs explicit information in view space (i.e. the angle between viewing direction and the light direction, which is relative to the camera in our data set).

Comparing the performance between the VNET* variants on the KIDNEY data, we can conclude that the additional channels and the trained TF of the VNETL16-4 and VNETL16-17 variants incrementally improve the models. Rendering the images in a latent color space of VNETL16-17 seems to provide a slight benefit for more complex tasks as in the KIDNEY data set, however VNET16-4 seems to produce better images on the SHADED training data. One reason might be that the additional parameters of VNET16-17 are harder to optimize in this case. On the KIDNEY data, the multiscale DVRNET architecture does not perform as well as the VNET* variants, missing both the visibility and distinct colorization in some cases. It seems that shifting the 3D upconvolutions of VNet to 2D upconvolutions in the design of our DVRNET is not as effective as the full VNet when complex 3D features are required.

From the results for the SHADED data set, it becomes apparent that DVRNET fares better at the view-dependent shading task. This can be explained with the fact that learning a view-dependent effect requires optimizing parameters that relate to view-spaces. Since all other architectures only have parameters that can perform reasoning in world space, Comparing qualitative and quantitative results, it seems that the perceptual scores we use in the analysis do not capture well the subtle illumination differences in this data set. SSIM shows these differences most clearly.

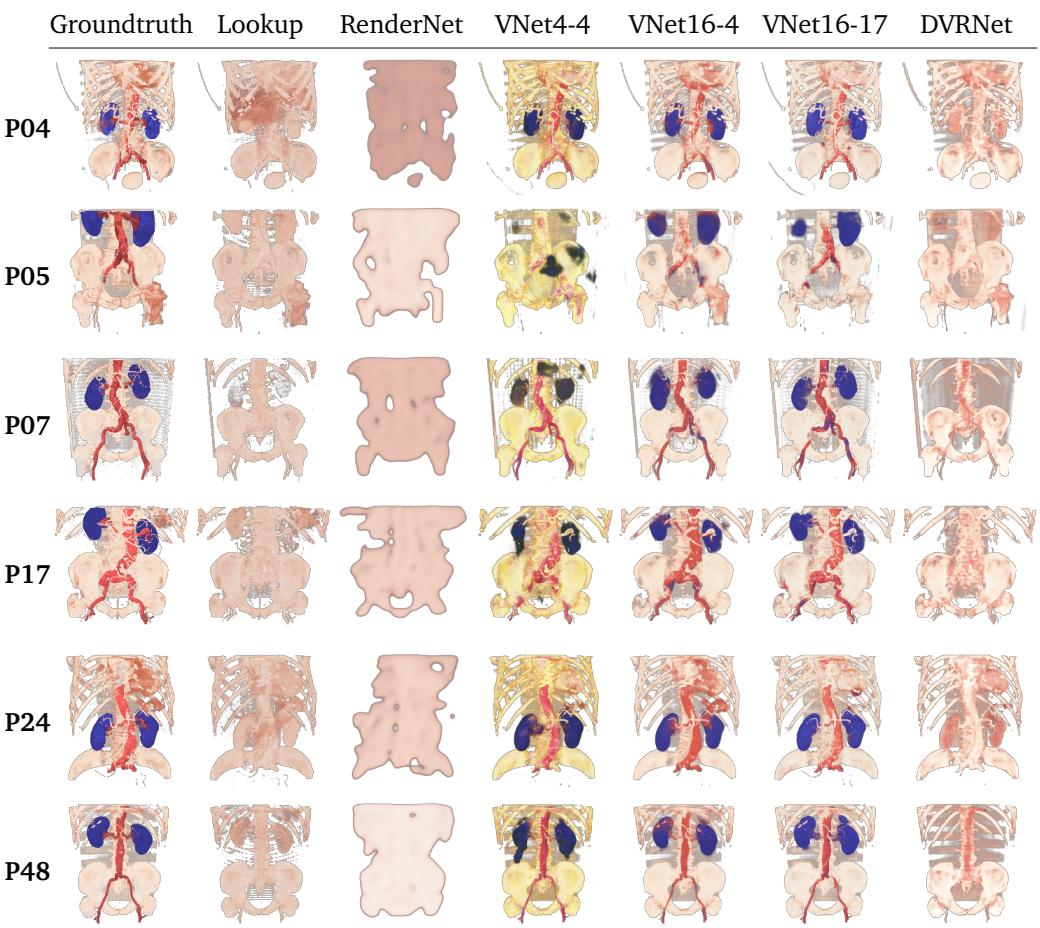


Fig. 10.13. Results for all our test patients on the KIDNEY data set. VNET16-4 and VNET16-17 perform best on this data set. DVRNET did not learn any colorization specific to the artery and kidneys.

Intuition: Learning Directional Illumination

Directional illumination can either be world-relative or view-relative, i.e. the light source moves with the camera. Intuitively, learning about a world-aligned directional light source can be performed by 3D convolutions: Considering for example Lambertian directional lighting along the X-axis, the illumination is can be directly computed from the volume gradient direction in X, for which a convolutional filter can even be defined analytically. However, from the view of a 3D CNN, trying to reason about view-dependent illumination is inconsistent because every training sample would provide different illumination. and the CNN does not have any knowledge about the view direction. It is therefore forced into computing the average illumination from all directions, which is equivalent a form of ambient occlusion. Deep learning-based ambient occlusion has been achieved before [58] in a more explicit fashion which likely outperforms our un-intentional approach. However, could be interesting to see if this restriction could be used in other applications to intentionally learn an average over view-dependent observations.

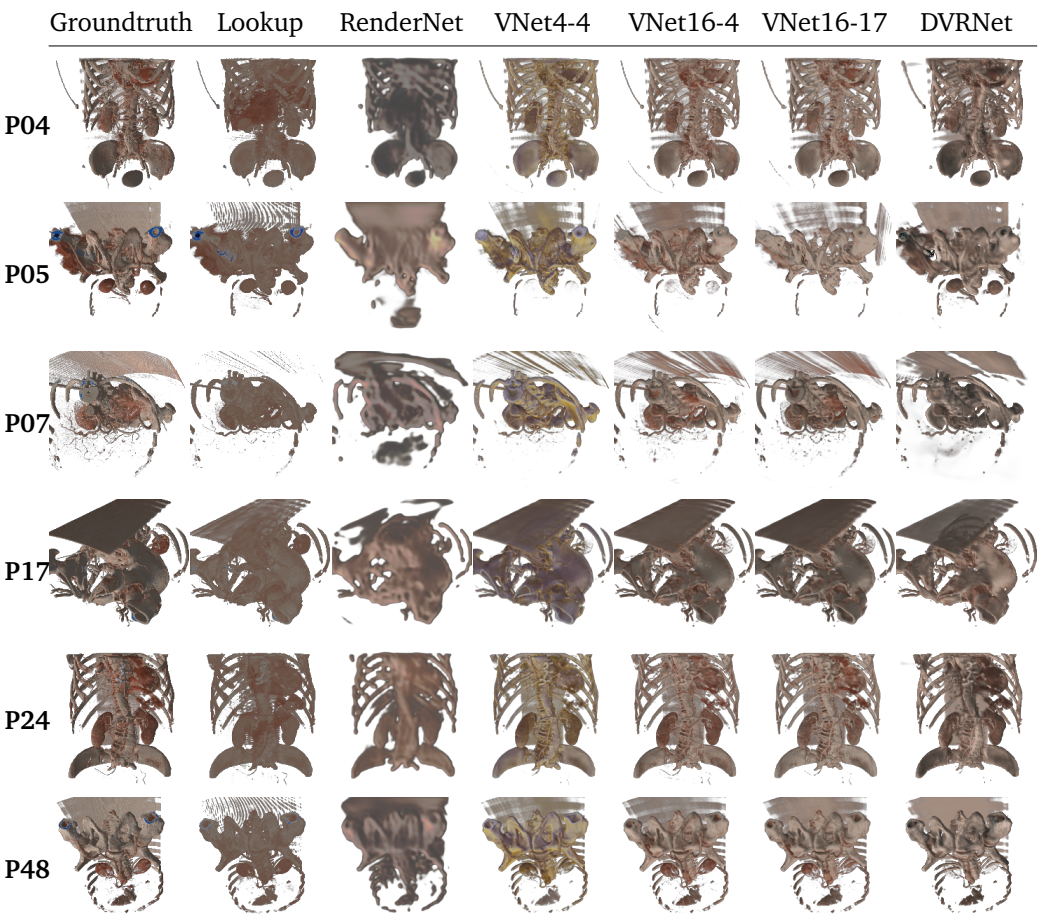


Fig. 10.14. Results for all test patients on our SHADED data set from a set of different viewing directions. DVRNet has learned the consistent illumination direction with respect to the camera. Other models only learned some static shading or the average luminance of the surface.

(a) Groundtruth (b) VNet16-17 (c) DVRNET

Fig. 10.15. Enlarged view of the results for **P48** in the SHADED dataset

Tab. 10.3. Results for the TF generalization task for the two data sets KIDNEY and SHADED. [1]SSIM was used as part of the loss function.

	KIDNEY			
	LPIPS ↓	FID ↓	SSIM ↑[1]	Train Time
Lookup	0.2627	**187.68**	**0.7060**	22m
RenderNet	0.5619	387.16	0.5368	3h 52m
VNet4-4	0.2683	280.26	0.6990	1d 2h 46m
VNetL16-4	0.2148	255.82	0.6936	1d 5h 19m
VNetL16-17	**0.2107**	245.45	0.6902	1d 5h 26m
DVRNet	0.2731	313.66	0.6993	7h 20m
	SHADED			
	LPIPS ↓	FID ↓	SSIM ↑[1]	Train Time
Lookup	0.2799	232.85	0.6765	22m
RenderNet	0.4799	262.71	0.6218	3h 53m
VNet4-4	0.2872	235.57	0.6640	1d 2h 52m
VNetL16-4	**0.2405**	**225.56**	0.6772	1d 5h 27m
VNetL16-17	0.2740	231.17	0.6561	1d 5h 32m
DVRNet	0.2485	227.85	**0.7337**	7h 23m

Notably, while RenderNet generally still produces blurry results, it seems to capture more illumination details from the SHADED images. This can be understood from its architecture, which transforms the volume to view space *before* 3D convolutions, which helps in understanding view-dependent effects. We consider this an interesting direction for further refinement of our own architectures.

10.5 Discussion

We have shown in our experiments that our novel DeepDVR successfully enables end-to-end training of neural rendering architectures for volume rendering and can produce neural rendering models that achieve the desired visual outcome without manual design of input features or TFs.

Our experiments have uncovered several characteristic differences between the presented architectures: The Lookup table provides a parametrization that remains manually adjustable through traditional user interfaces and fast training times. When only using intensity as an input feature, it is ultimately limited in its ability to discern different structures. Nonetheless, training a Lookup TF is useful in applications where the integration of a full neural renderer is not possible. In those cases, the lookup table can simply be transferred to existing DVR implementations. The training approach can also be readily extended to 2D TFs, where manual specification is less straightforward.

RenderNet overall is clearly outperformed by the other deep architectures we have investigated which demonstrates the effectiveness of our dedicated DVR unit. VNET16-17 seems to perform similar to VNET16-4 on the TF generalization task yet it provides better results on the manually adapted reference images. DVRNET overall performed best on the manually adapted images and produced comparable renderings for the generalization tasks while requiring significantly

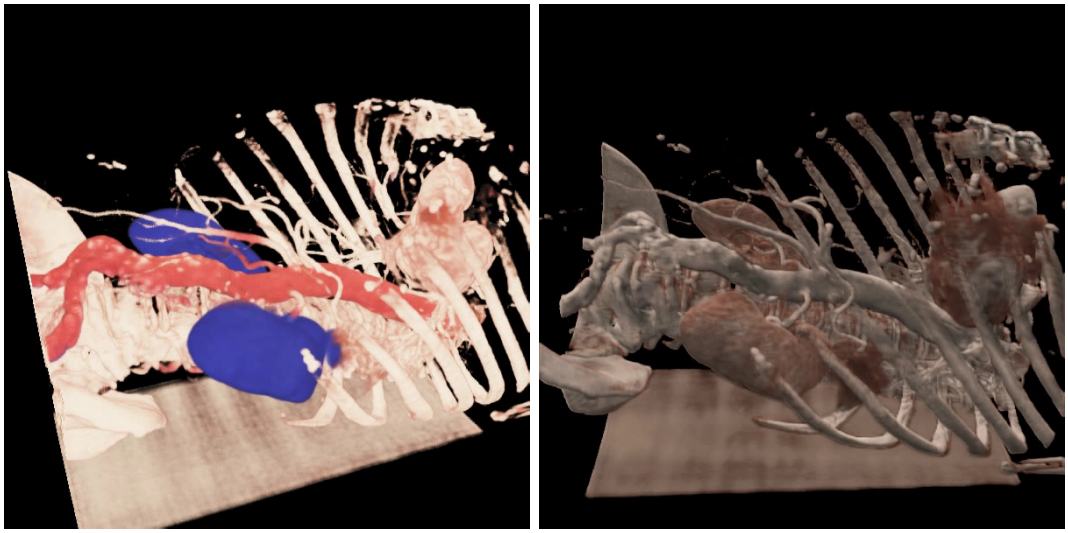


Fig. 10.16. Our models can generate high resolution output images even though being trained only on low resolution data training data. Left: VNetL16-17 trained on the KIDNEY data set. Right: DVRNET trained on the SHADED data set.

less time for training than the VNet variants. The specific architecture choices for DVRNET, involving both 3D and 2D convolutions, allow to capture both 3D spatial semantics and view-dependent aspects of the training images. However, the reduced number of input convolutions in the volume encoder make it less powerful than the full VNet-based encoders.

We want to highlight that the resampling step in the ray casting within the model architectures effectively decouples the input volume resolution from the output image resolution. Because we have chosen fully convolutional layers in the 3D and 2D portions of the architectures, our models can flexibly process arbitrary volume resolutions and produce arbitrary image resolutions. The images in Figure 10.16 were rendered with the the two best-performing model trained in section 10.4.3 but with an increased output resolution of 600×600 as opposed to 200×200 during training.

Limitations and Future Work It is important to acknowledge that our proposed deep architectures prevent the direct adaptation in cases where the rendering is not acceptable for a new volume. While removing manual interaction was the dedicated goal of our method, this is relevant until models are good enough that they "just work", so this argument merits consideration. We could for example imagine a hybrid combination of learned 2D lookup table and learned features, where the lookup table remains editable with conventional interfaces. Perceptual loss functions[100, 236] or GANs could also be used for a final patient-specific optimization step of a pretrained model to account for per-patient variations outside of the training examples.

The high training time of the deep variants currently limits the range of applications to where precomputation can be done offline. However, the modular architecture and explicit modeling of color and opacity functions in our models enable defining selected presets or even partial retraining. Interactive refinement based on user feedback is an interesting future research direction, where we expect the explicit modeling of the DVR stages to play a crucial role. In

this work we did not focus on optimizations of the differentiable rendering algorithm itself. At test time, our current implementation of DVRNet takes approximately 370ms per frame, whereas the VNet* variants take 1580-2008ms. An explicit implementation of the forward pass directly using shaders or GPU computing APIs together with caching the output from the volume encoder $\mathcal{E}(I)$ can provide significant speedups that will likely enable interactive visualization. Further research is also required to reduce the memory footprint during training which will enable training on larger input volumes and output image sizes.

Beyond direct applications in rendering, we believe that the presented DeepDVR approach can also be used to train 3D segmentation networks from annotations in view space. Our VNET4-4 model is essentially a 3D semantic segmentation model optimized by a visual loss function by projecting the 2D annotations into 3D. The experiments demonstrate that differentiable volume rendering in the training loop can be used to train semantic segmentation models from scratch only from simple 2D annotations.

The process of accumulating features along a ray from front to back is essentially an application of sequence processing and the alpha blending can be seen as a form of recurrent neural network. From this perspective, using RNN units like LSTM or GRU units could in the future replace the accumulation function $\mathcal{A}$ for learning more complex blending operations. This could be used to learn advanced implicit ray accumulation functions and produce learned "smart" compositing and illustrative rendering.

10.6 Conclusion

In this chapter, we explored the possibilities and applications of neural networks modeling the DVR pipeline. We show that with this approach, deep neural networks are able to learn the hidden semantic of what structures the user *wants* to see purely from their indications in image space. This includes complex relations where the system learned to discriminate between larger and smaller vessels to only highlight the artery. This act of inferring the 3D semantic that the user is looking for, and adapting the visualization accordingly, will be an immensely powerful tool to enable future intuitive user interaction with complex or high-dimensional data.

Our experiments have shown that representing a learned 1D transfer function using MLPs has little benefit over the direct representation as a lookup table, however the application to multidimensional feature spaces (i.e. in our advanced architectures) show that MLPs are still an important tool. We have also presented several architectures that are motivated by the careful modeling of the generalized volume rendering algorithm we have introduced. We have highlighted two practically relevant applications for DeepDVR: TF design by manual annotation in image space, and learning a generalized, robust rendering model from a set of individually fine-tuned renderings. Furthermore, we have introduced *stepsize annealing*, a method to accelerate training of ray casting models by starting at a high rendering step size in early epochs, decreasing over time to achieve similar performance as when training with a constant small step size.

We consider the DeepDVR approach as an *architectural blueprint* which can be combined with any 2D and 3D image-to-image architectures. Building a DeepDVR architecture that supports both good classification in volume space and good image-space reasoning could be as simple as combining a full VNet-based volume encoder with a full U-Net in the decoder. However, the biggest challenge in the future will be to find lightweight architectures and algorithmic advances that enable interactive visualization and online retraining.

The high computational demands required for the models presented in this chapter make their direct application to real-time intraoperative visalization of OCT data challenging. With further computational optimization of the inference and rendering, it will likely be feasible to use for visualization of static OCT volumes. In the context of our requirements, this work is an isolated step towards the development of specialized (**III.**) and more semantically aware visualizations (**IV.**) while intentionally ignoring the other dimensions to explore novel technologies. Future improvements could make the approach more performant for example by considering only lower-dimensional, more explicit feature spaces and only learning the complex transfer functions from target images.

Part IV

Conclusion

11 Discussion and Conclusion

11.1 Summary

Contents

In this dissertation, we have addressed several issues surrounding intraoperative visualization in the context of retinal microsurgery. Our presented solutions bridge the gap between OCT imaging technology and modern visualization assisted by deep learning.

The vision of the visualization-guided robotic surgery environment we have presented in section II was the primary driver for the research agenda underlying the presented methodology. We have outlined how a such future system can integrate the multimodal information available from OCT, microscope and robot and combine it into a dynamic visualization and control system. By focusing on building early *proof of concepts* of these parts of the system, we could identify the principle advances necessary towards achieving it. We will now revisit those requirements to put our contributions into a larger context.

I. High-performance OCT Imaging For a system to enable real-time feedback on intraoperative maneuvers, both temporal and spatial image resolution have to be sufficiently high. The advancements in SS-OCT technology of the recent years provide engines with A-scan rates that are sufficiently to achieve real-time volumetric imaging and even put OCT-only microscopy within reach. However, the high technical expertise required in the complex setup as well as the high costs associated with the necessary components make such setups prohibitive for many research labs. The additional effort of integrating with a surgical microscope means that only very few research centers currently have the ability to perform (clinical) studies with high performance iOCT.

II. Efficient Processing The vast amounts of multimodal data provided by the OCT and stereo microscope need to be processed at speed in an intraoperative setting, imposing computational bounds on any algorithmic processing. In this dissertation, we have employed two principle approaches to manage the computational complexity of processing 4D OCT:

The instrument tracking method presented in section 7 models the dynamic motion of instrument and OCT explicitly. It relies solely on processing B-scans and is therefore amenable to sparse sampling of the dense volumetric information. This can be exploited to control the computational budget that is used for the tracking algorithm, for example by only processing

a fixed subset of the B-scans of a volume. Alternatively, one could implement a *best effort processing* scheme where a fixed computational budget (e.g. a single thread) is alotted for tracking, always processing only the latest B-scan available. The second approach we use to reduce computation in our real-time visulization (section 9) is to leverage axial projections. These images that are fast to compute and exploit the directional nature of the imaging modality as well as the fact that the retinal tissue is mostly perpendicular to the A-scan direction. This reduces the processing domain to a 2D problem for each volume and allows us to apply well-researched computer vision algorithms to perform temporal alignment as well as semantic segmentation.

While GPU and CPU speeds and bandwidths are ever increasing and some of the bottlenecks might be resolved by simply using more powerful hardware, the imaging rate of A-scans as well as the processing demands of more complex algorithms are progressing in parallel. Therefore, schemes for efficient processing will continue to be important.

III. Specialized Visualization Conventional 3D imaging modalities like MRI or CT have had decades of research dedicated to them to develop and improve visualizations. Owing to its relative novelty, 3D and 4D OCT has only received little attention in this regard despite its uniquely interesting challenges.

We have presented two complementary visualization modes directly developed in response to the clinical need for better visualization of 3D OCT data in section 8: LA-MIP is a projective visualization designed to clearly show the instrument below the retinal surface during subretinal injections. The layer-aware DVR method takes into account the layered structure of the retinal tissue and uses a specifically chosen shading model to enhance the perception of the retinal surface structure. In our follow up work on 4D visualization (section 9), we have not only shown how this concept can be adapted to SS OCT but also extended the real-time rendering to enhance instrument visibility, compensate shadowing artefacts and improve general image quality.

In search for a more general solution for volume visualization without the need for explicit labels or transfer functions, we have developed Deep Direct Volume Rendering (section 10), an approach that puts the direct volume rendering algorithm directly into the trainable deep learning pipeline. Even though the performance requirements as of now are prohibitive, we think that this can be the next step towards a better semantic visualizations. With other research groups already embracing research into differentiable DVR, improvements on the computational aspects of inference and training are already being published and could in the future even lead to adaptive visualizations that continuously optimize the viewpoint, transfer function or other rendering parameters based on the latest iOCT data.

IV. Semantic Awareness Building a system that makes effective use of the available data needs to be aware of the semantic content of the surgical scene in order to provide intelligent assitance to the surgeon.

Several components of our proposed methods push in this direction: our 5DOF instrument tracking algorithm already extracts some knowledge about instrument and layer positioning in the OCT data. The visualization methods presented in Chapters 8 and 9 are based on a

notion of layer positioning and instrument labeling to provide improved guidance. However, current deep learning methodology for semantic segmentation offers much richer possibilities and works in workflow recognition could be especially interesting to automatically adapt visualizations or robotic modes. As robotic control transitions towards more autonomy, this will also require a deeper and more wholistic understanding of the surgical scene by the robotic system, which could in turn be mirrored back to the visualization for an observer.

DeepDVR sidesteps an explicit semantic segmentation in favor of learning the semantics implicitly from example images. While this makes it hard for other systems to directly use this semantic knowledge that is internal to the rendering model, the generalized way of formulating the rendering as apart of a deep learning based pipeline allows for a more natural integration of the existing semantic information (i.e. the layer segmentations or instrument labels) into the rendering. It also creates the potential for end-to-end training of label segmentation and visualization in one go, which could have interesting benefits.

V. Robotic Automation With our robotic platform being restricted to direct control by surgeon via a joystick or space mouse, missing advanced automation was one of the greatest pain points identified in our early experiments with system integration. The robotic community is actively researching better ways to improve control for general endoscopic robots as well as robots for microsurgery. Smart virtual fixtures and adaptive motion scaling for example could greatly improve handling of the robot. When combining this with semantic information extracted from OCT, safer control mechanisms could be created e.g. by reducing maximum speed when close to the target or risk areas. Integration of such approaches is complex due to the bespoke nature of the robot control software used, and has so far not been achieved for the robotic platform used in our experiments.

VI. End-to-End Integration The integration of the three principle platforms for imaging, robotics and visualization is a demanding problem that requires careful planning. Heterogeneous platforms in combination with the high performance and low latency requirements of the use case introduce further complexities into the overall system design. Computational resources and data pathways need to be carefully controlled to ensure stable system performance.

Based on our experience with the prototypes developed in section II, we firmly believe that a wholistic approach not only in terms of communication architecture but also algorithmically is necessary to achieve the envisioned system. We have pursued this philosophy in the development of our methodology, where we have always tried to use and reuse information that results from intermediate steps in the algorithms. This is most apparent in our pipeline for the processing-aware rendering, where the projection maps are used and reused not only for instrument segmentation but also for temporal registration and during rendering. The same projection maps and instrument label map could even be used in an adapted form of our 5DOF tracking algorithm to add tracking with low overhead to the system.

However, the full integration of all required technology will be one of the greatest hurdles towards realizing an integrated system. Even though open data exchange protocols like ROS or OpenIGTL offer standardized ways to exchange a variety of data types, these do not scale well with higher bandwidth. Ideally, processing of the imaging data is performed with a

multi-GPU system to manage the bandwidth and latency requirements, however this will require significant architectural design and implementation efforts to achieve.

11.2 Outlook

Working with an imaging modality that is undergoing a fundamental technological transition towards increasing imaging speeds by an order of magnitude is an exceptional experience that has driven much of the vision we have presented. Parallel advances in robotics and machine learning have generated a constant stream of inspiration and will continue to provide novel opportunities for advances in this area. In particular, progress in natural language interfaces, semantic scene understanding and autonomous agents will allow future systems to react in more nuanced ways to a surgeon's requirements and even perform autonomous tasks.

The realities of working with such novel technologies have often made it necessary for us to find the right abstractions for research in order to avoid solving the problems of previous generation devices and instead have a lasting impact on future technology. However, as technologies mature and best practices and communication standards evolve, integration of these complex systems will become more manageable and allow creating connected systems with less effort. It will also enable further clinical evaluation of the work we have presented in this dissertation, which is a critical next step towards clinical translation of the methods.

Digital microscopes and commercial robotic systems for microsurgery are already changing the look of modern eye surgery theaters. In developing the vision of the visualization-guided robotic surgery environment we have presented in section II, we have shown a path towards a future surgical theater where all these systems are interconnected and centered around a smart visualization system. The more radical transformation towards the VR-integrated robotic surgical cockpit might still be a long way from the current manual surgery performed through the binocular microscope. Yet as both mixed reality and robotic technology evolve with time and pervade our daily lives, we expect that the adoption of such technologies into the OR will become commonplace in some form or another: Both technologies are an opportunity to enhance a surgeon's capabilities beyond human limits of perception or manipulation and both will be required for the highly precise interventions of the future.

Working together with partners from industry as well as several technical and clinical research groups has been instrumental in much of the work we have presented here. On a personal note, I firmly believe that this kind of interdepartmental and interdisciplinary research, while not without its own challenges, is the best way towards impactful translational research, in particular in the challenging domain of eye surgery. While it seems sometimes easier (and it often is truly necessary) to break up a system and consider research problems in isolation, working towards a grand vision also requires sometimes stepping back and considering the big picture in order to avoid ending up with a collection of only locally optimal solutions. I hope that the vision outlined in this dissertation will inspire future research in the direction of VR-assisted robotic microsurgery and that the methodologies provided will prove to be valuable puzzle pieces that will some day be part of a more complete picture.

Part V

Appendix

List of Authored and Co-authored Publications

A

Authored

1. Alejandro Martín-Gomez[†], **Jakob Weiss**[†], Andreas Keller, Ulrich Eck, Daniel Roth, Nassir Navab. "The Impact of Focus and Context Visualization Techniques on Depth Perception in Optical See-Through Head-Mounted Displays". *IEEE Transactions on Visualization and Computer Graphics, to appear, 2021. [†]Authors contributed equally.*

2. **Jakob Weiss**, Nassir Navab. "Deep Direct Volume Rendering". *arXiv preprint, 2021*

3. Gloria Zörnack[†], **Jakob Weiss**[†], Georg Schummers, Ulrich Eck, Nassir Navab. "Evaluating surface visualization methods in semi-transparent volume rendering in virtual reality". *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, Vol. 9, Issue 4, 2021. [†]Authors contributed equally.*

4. **Jakob Weiss**, Michael Sommersperger, M. Ali Nasseri, Abouzar Eslami, Ulrich Eck, Nassir Navab. "Processing-Aware Real-Time Rendering for Optimized Tissue Visualization in Intraoperative 4D OCT". *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI), 2020, Lima, PE*

5. **Jakob Weiss**, Ulrich Eck, M. Ali Nasseri, Mathias Maier, Abouzar Eslami, Nassir Navab. "Layer-Aware iOCT Volume Rendering for Retinal Surgery". *Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM), pp.123-127, 2019, Brno, CZ*

6. **Jakob Weiss**, Nicola Rieke, M. Ali Nasseri, Mathias Maier, Chris P. Lohmann, Nassir Navab, Abouzar Eslami. "Injection Assistance via Surgical Needle Guidance using Microscope Integrated OCT (MI OCT)". *Annual Meeting of the Association for Research in Vision and Ophthalmology (ARVO), 2018, Honolulu, US*

7. **Jakob Weiss**, Nicola Rieke, M. Ali Nasseri, Mathias Maier, Abouzar Eslami, Nassir Navab. "Fast 5DOF Instrument Tracking in iOCT". *International Journal of Computer Assisted Radiology and Surgery, Vol. 13 Issue 2, pp.787-796, 2018*

Co-Authored

1. Philipp Matten, Anja Britten, Michael Niederleithner, **Jakob Weiss**, Hessam Roodaki, Benjamin Sorg, Nancy Hecker-Denschlag, Wolfgang Drexler, Rainer Leitgeb, Tilman

Schmoll. "MHz SS-OCT–from biometry to live volumetric imaging". *Investigative Ophthalmology & Visual Science Vol. 62 Issue 11, pp.43-43, 2021*

2. Anja Britten, Philipp Matten, Michael Niederleithner, **Jakob Weiss**, Hessam Roodaki, Benjamin Sorg, Nancy Hecker-Denschlag, Wolfgang Drexler, Rainer A Leitgeb, Tilman Schmoll. "A multipurpose SS-OCT engine – from axial eye length measurements to 4D OCT". *Investigative Ophthalmology & Visual Science Vol. 62 Issue 11, pp.2502-2502, 2021*

3. Anja Britten, Philipp Matten, Michael Niederleithner, **Jakob Weiss**, Wolfgang Drexler, Rainer A Leitgeb, Tilman Schmoll. "Versatile MEMS-VCSEL SS-OCT engine: full eye biometry to 4D imaging at MHz A-scan rates". *Optical Coherence Tomography and Coherence Domain Optical Methods in Biomedicine XXV, 2021*

4. Michael Sommersperger, **Jakob Weiss**, M Ali Nasseri, Peter Gehlbach, Iulian Iordachita, Nassir Navab. "Real-time tool to layer distance estimation for robotic subretinal injection using intraoperative 4D OCT". *Biomedical Optics Express Vol. 12 Issue 2, 2021*

5. Arianne Tran, **Jakob Weiss**, Shadi Albarqouni, Shahrooz Faghi Roohi, Nassir Navab. "Retinal Layer Segmentation Reformulated as OCT Language Processing". *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI), 2020, Lima, PE*

6. Mingchuan Zhou, Xijia Wang, **Jakob Weiss**, Abouzar Eslami, Kai Huang, Mathias Maier, Chris P Lohmann, Nassir Navab, Alois Knoll, M Ali Nasseri. "Needle localization for robot-assisted subretinal injection based on deep learning". *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) pp.8727-8732, 2019*

7. Mingchuan Zhou, Mahdi Hamad, **Jakob Weiss**, Abouzar Eslami, Kai Huang, Mathias Maier, Chris P. Lohmann, Nassir Navab, Alois Knoll, M. Ali Nasseri. "Towards robotic eye surgery: Marker-free, online hand-eye calibration using optical coherence tomography images". *IEEE Robotics and Automation Letters Vol. 3 Issue 4, pp.3944-3951, 2018*

8. Florian Reichl, **Jakob Weiss**, Rüdiger Westermann. "Memory-Efficient Interactive Online Reconstruction From Depth Image Streams". *Computer Graphics Forum Vol. 35 Issue 8, pp.108-119, 2016*

9. Christian Schulte zu Berge, **Jakob Weiss**, and Nassir Navab. "Schematic Electrode Map for Navigation in Neuro Data Sets". *Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM), 2015, Chester, UK*

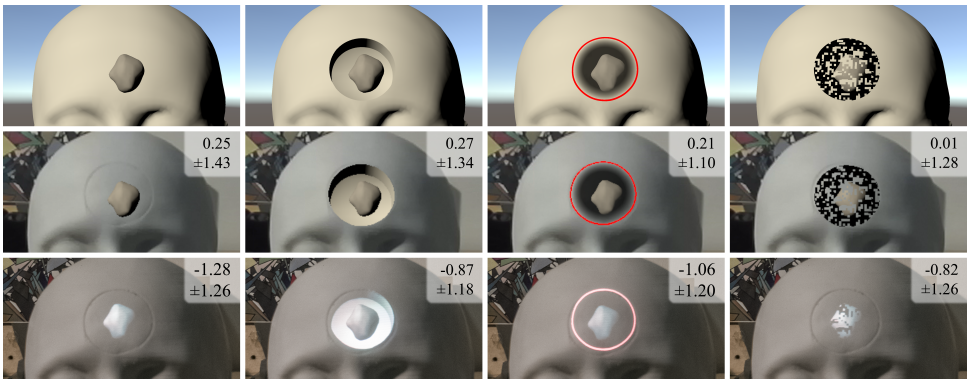
10. Amit Shah, Oliver Zettinig, Tobias Maurer, Cristina Precup, Christian Schulte zu Berge, **Jakob Weiss**, Benjamin Frisch, Nassir Navab. "An open source multimodal image-guided prostate biopsy framework". *Workshop on Clinical Image-Based Procedures, 2014*

B

Abstracts of Publications not Discussed in this Thesis

The Impact of Focus and Context Visualization Techniques on Depth Perception in Optical See-Through Head-Mounted Displays

Alejandro Martín-Gomez†, Jakob Weiss†, Andreas Keller,
Ulrich Eck, Daniel Roth, Nassir Navab
†*Authors contributed equally.*



Base implementations of focus and context visualization techniques (*top row*) and their appearance in video- (*mid row*), and optical- (*bottom row*) see-through head-mounted displays. From left to right: *Baseline* overlay without contextual layer, *Virtual Window*, *Contextual Anatomical Mimesis*, and *Virtual Mask*. Mean and standard deviation of corresponding alignment errors of study 1 are presented in centimeters. The OST images are captured using a smartphone camera placed at the eye position. Contrast and brightness have been adjusted for a faithful impression of the overlay as observed by the user.

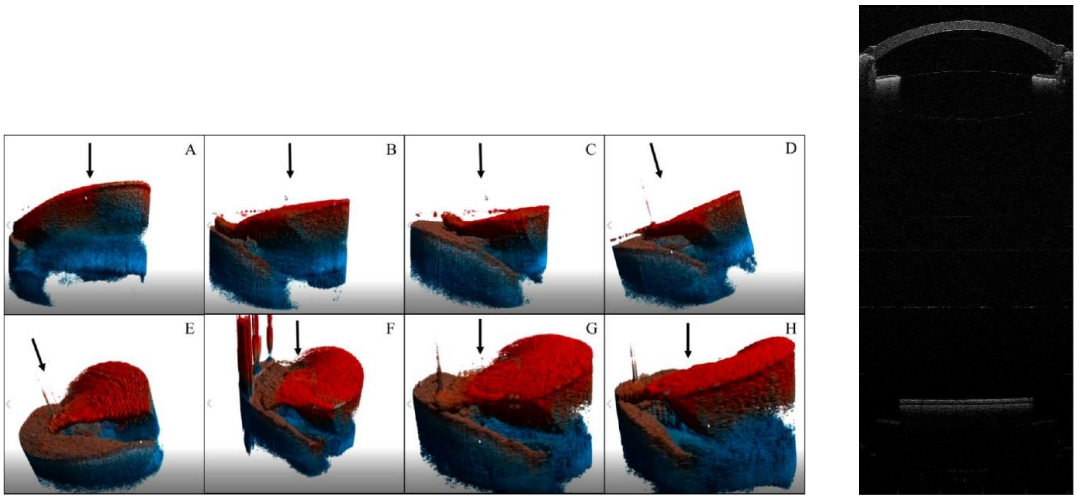
Estimating the depth of virtual content has proven to be a challenging task in Augmented Reality (AR) applications. Existing studies have shown that the visual system uses multiple depth cues to infer the distance of objects, occlusion being one of the most important ones. Generating appropriate occlusions becomes particularly important for AR applications that require the visualization of augmented objects placed below a real surface. Examples of these applications are medical scenarios in which anatomical information needs to be observed within the patients body. In this regard, existing works have proposed several focus and context (F+C) approaches to aid users in visualizing this content using Video See-Through (VST) Head-Mounted Displays (HMDs). However, the implementation of these approaches in Optical See-Through (OST) HMDs remains an open question due to the additive characteristics of the display technology.

In this paper, we, for the first time, design and conduct a user study that compares depth estimation between VST and OST HMDs using existing in-situ visualization methods. Our results show that these visualizations cannot be directly transferred to OST displays without increasing error in depth perception tasks. To tackle this gap, we perform a structured decomposition of the visual properties of AR F+C methods to find best-performing combinations. We propose the use of chromatic shadows and hatching approaches transferred from computer graphics. In a second study, we perform a factorized analysis of these combinations, showing that varying the shading type and using colored shadows can lead to better depth estimation when using OST HMDs.

IEEE Transactions on Visualization and Computer Graphics (2021)

MHz SS-OCT–from biometry to live volumetric imaging

Philipp Matten, Anja Britten, Michael Niederleithner, Jakob Weiss,
Hessam Roodaki, Benjamin Sorg, Nancy Hecker-Denschlag,
Wolfgang Drexler, Rainer Leitgeb, Tilman Schmoll



Left: Image sequence of live imaging of a porcine cornea at 17vol/s. Right: Full-length ocular biometry scan.

Purpose: For various applications during ophthalmic surgery spectral-domain optical coherence tomography (SD-OCT) is limited through its low A-scan rate and imaging depth. We present a versatile swept-source OCT (SS-OCT) engine, which addresses a large collection of use cases, ranging from axial eye length measurements to live volumetric visualizations at MHz A-scan rates.

Methods: We developed a flexible SS-OCT engine and an add-on module to couple its sample arm to an ophthalmic surgical microscope. This engine includes a 1060nm tunable MEMS-VCSEL whose sweep repetition rate can be alternated between 100kHz, 400kHz or 1MHz. To increase the effective A-scan rate at the cost of axial resolution, we scanned at twice the speed and mathematically divided each sweep into two halves. The 100kHz mode was used for high resolution B-scan and axial eye length measurements with an axial resolution of 6.3mm and an imaging depth of 29mm in tissue. For 4D live imaging, effective A-scan rates of 800kHz and 2MHz allowed us to realize fields of view (FOVs) of 3.1-15.7mm with imaging depths of 4.3-10.5mm and an axial resolution of 12.6mm in tissue. We imaged anterior segment and retina mimicking phantom eyes, as well as ex vivo porcine eyes. All data was processed and rendered live.

Results: Using the same instrument, we acquired full eye scans, anterior and posterior segment B-scans, as well as 4D-OCT scans with volume rates of up to 17vol/s. Fig.1 shows an ocular biometry scan of a test eye captured at an A-scan rate of 100kHz. Sampling such enormous depths of 29mm in real-time allows for ocular distance measurements or solid state z-tracking. In Fig.2A-H an image sequence of a 17vol/s live rendered volume series can be seen. It visualizes an incision of a porcine cornea displayed at different viewing angles and zooms. An

A-scan rate of 1MHz and axial resolution of $6.3\,\mu\mathrm{m}$ resulted in a FOV of 3.1mm. To further enhance depth perception, depth is color-encoded from red (top) to blue (bottom) and the z-direction is indicated by an arrow.

Conclusions: We demonstrated SS-OCTs' potential to address multiple ophthalmic imaging applications with a single device. 4D OCT can be used for enhancing depth perception and visualizing sub-surface structures during surgery. Imaging at lower rates enables full eye OCT for high resolution B-scan imaging and biometry.

Investigative Ophthalmology & Visual Science Vol. 62 Issue 11 (2021)

Versatile MEMS-VCSEL SS-OCT engine: full eye biometry to 4D imaging at MHz A-scan rates

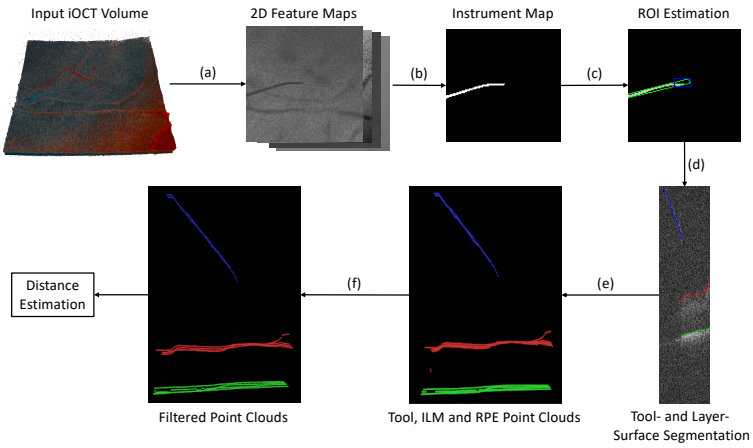
Anja Britten, Philipp Matten, Michael Niederleithner, Jakob Weiss,
Wolfgang Drexler, Rainer A Leitgeb, Tilman Schmoll

We present a flexible OCT engine for acquiring full eye-length, anterior and posterior segment B-scans, as well as 4D live volumes with an effective A-scan rate of up to 2MHz. It is enabled by a MEMS tunable VCSEL with flexible A-scan rates, broad spectral bandwidth and a long instantaneous coherence length. Our GPU based, custom reconstruction and rendering software is able to process and display live volume series at rates of up to 17 volumes per second. We show B-scans and volume series of model eyes.

Optical Coherence Tomography and Coherence Domain Optical Methods in Biomedicine XXV (2021 Conference Presentation)

Real-time tool to layer distance estimation for robotic subretinal injection using intraoperative 4D OCT

Michael Sommersperger, Jakob Weiss, M. Ali Nasseri, Peter Gehlbach, Iulian Iordachita, Nassir Navab



An overview of our pipeline. (a) A set of 2D feature maps is generated from the newly acquired volume. (b) Instrument segmentation based on the projection images yields a binary tool map. (c) Estimation of a small ROI around the needle tip (indicated by the blue rectangle). (d) The tool and retinal layer boundaries of the selected B-scans within the ROI are segmented. (e) The three point clouds corresponding to instrument, ILM and RPE surface boundaries are generated from the segmentation maps. The shadowed retinal layers are reconstructed considering the surrounding anatomy. (f) Noise is removed from the point clouds by applying Euclidean clustering. The distance between instrument and retinal layers is finally obtained from the resulting point clouds.

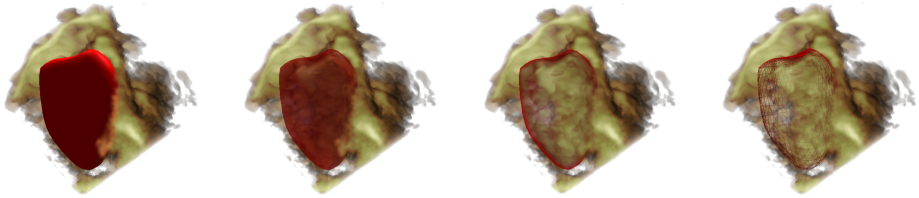
The emergence of robotics could enable ophthalmic microsurgical procedures that were previously not feasible due to the precision limits of manual delivery, for example, targeted subretinal injection. Determining the distance between the needle tip, the internal limiting membrane (ILM), and the retinal pigment epithelium (RPE) both precisely and reproducibly is required for safe and successful robotic retinal interventions. Recent advances in intraoperative optical coherence tomography (iOCT) have opened the path for 4D image-guided surgery by providing near video-rate imaging with micron-level resolution to visualize retinal structures, surgical instruments, and tool-tissue interactions. In this work, we present a novel pipeline to precisely estimate the distance between the injection needle and the surface boundaries of two retinal layers, the ILM and the RPE, from iOCT volumes. To achieve high computational efficiency, we reduce the analysis to the relevant area around the needle tip. We employ a convolutional neural network (CNN) to segment the tool surface, as well as the retinal layer boundaries from selected iOCT B-scans within this tip area. This results in the generation and processing of 3D surface point clouds for the tool, ILM and RPE from the B-scan segmentation maps, which in turn allows the estimation of the minimum distance between the resulting

tool and layer point clouds. The proposed method is evaluated on iOCT volumes from ex-vivo porcine eyes and achieves an average error of $9.24\,\mu\mathrm{m}$ and $8.61\,\mu\mathrm{m}$ measuring the distance from the needle tip to the ILM and the RPE, respectively. The results demonstrate that this approach is robust to the high levels of noise present in iOCT B-scans and is suitable for the interventional use case by providing distance feedback at an average update rate of 15.66 Hz.

Biomedical Optics Express Vol. 12 Issue 2 (2021)

Evaluating surface visualization methods in semi-transparent volume rendering in virtual reality

Gloria Zörnack†, Jakob Weiss†, Georg Schummers, Ulrich Eck, Nassir Navab
†*Authors contributed equally.*



Mesh geometry with ultrasound volume rendering. Evaluated visualization techniques (from left to right): OPAQUE, SEMI-TRANSPARENT, SILHOUETTE, WIREFRAME.

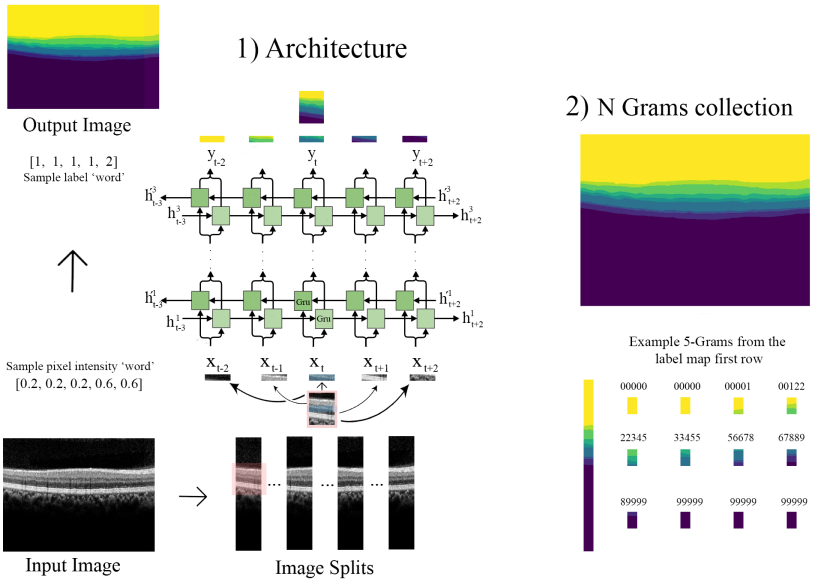
Perceptual visualization of semi-transparent structures in volumetric datasets ischallenging due to its inherent visual complexity. This is however of primary impor-tance in medical image visualization where volume rendering of 3D data is a commontechnique. While volume rendering techniques are a well-researched area, perceptionof volume-rendered voxel data in combination with semi-transparent mesh data re-mains an underexplored topic. As virtual reality (VR) is increasingly employed forvolume data inspection in the medical domain, it becomes important to understandhow different mesh visualizations affect performance in medical tasks in this context.

In this work, we compare visualization techniques when combining geometry withdirect volume rendering in immersive VR, where stereoscopic vision and motionparallax provide additional depth cues.In this work, we investigate how different surface transparency modes affect taskperformance in a VR setup. For two medical image analysis tasks, we conduct auser study (n=23) to analyze the impact various mesh rendering methods have ontask outcome and subjective preference. Our evaluation indicates that user perfor-mance when using wireframe rendering varies greatly between tasks while constantopacity and silhouette provide a stable benefit. Overall, the transparent visualiza-tions led to improved precision (lower error rate, lower inter-observer variability)and users were more confident about the outcome of their image analysis task, whilethe efficiency of each visualization method was task-dependent. Our results demon-strate how semi-transparent visualization will improve visual analysis tasks in VRfor medical applications.

Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, Vol. 9, Issue 4 (2021)

Retinal Layer Segmentation Reformulated as OCT Language Processing

Arianne Tran, Jakob Weiss, Shadi Albarqouni, Shahrooz Faghi Roohi, Nassir Navab



(1) Shows the architecture and the reformulation of the image into sequences (2) Displays an example of N-Gram gathering along a row, for visualization's sake the row has been widened in this figure to cover several columns

In the medical field, semantic segmentation has recently been dominated by deep-learning based image processing methods. Convolutional Neural Network approaches analyze image patches, draw complex features and latent representations and take advantage of these to label image pixels and voxels. In this paper, we investigate the usefulness of Recurrent Neural Network (RNN) for segmentation of OCT images, in which the intensity of elements of each A-mode depend on the path projected light takes through anatomical tissues to reach that point. The idea of this work is to reformulate this sequential voxel labeling/segmentation problem as language processing. Instead of treating images as patches, we regard them as a set of pixel column sequences and thus tackle the task of image segmentation, in this case pixel sequence labeling, as a natural language processing alike problem. Anatomical consistency, i.e. expected sequence of voxels representing retinal layers of eye's anatomy along each OCT ray, serves as a fixed and learnable grammar. We show the effectiveness of this approach on a layer segmentation task for retinal Optical Coherence Tomography (OCT) data. Due to the inherent directionality of the modality, certain properties and artifacts such as varying signal strength and shadowing form a consistent pattern along increasing imaging depth. The retinal layer structure lends itself to our approach due to the fixed order of layers along the imaging direction. We investigate the influence of different model choices including simple RNNS, LSTMs and GRU structures on the outcome of this layer segmentation approach. Experimental results show that the potential of this idea that is on par with state of the art works while being flexible to changes in the data structure.

Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI), (2020)

Needle localization for robot-assisted subretinal injection based on deep learning

Mingchuan Zhou, Xijia Wang, Jakob Weiss, Abouzar Eslami, Kai Huang,
Mathias Maier, Chris P Lohmann, Nassir Navab, Alois Knoll, M Ali Nasseri

Subretinal injection is known to be a complicated task for ophthalmologists to perform, the main sources of difficulties are the fine anatomy of the retina, insufficient visual feedback, and high surgical precision. Image guided robot-assisted surgery is one of the promising solutions that bring significant surgical enhancement in treatment outcome and reduces the physical limitations of human surgeons. In this paper, we demonstrate a robust framework for needle detection and localization in subretinal injection using microscope-integrated Optical Coherence Tomography (MI-OCT) based on deep learning. The proposed method consists of two main steps: a) the preprocessing of OCT volumetric images; b) needle localization in the processed images. The first step is to coarsely localize the needle position based on the needle information above the retinal surface and crop the original image into a small region of interest (ROI). Afterward, the cropped small image is fed into a well trained network for detection and localization of the needle segment. The entire framework is extensively validated in ex-vivo pig eye experiments with robotic subretinal injection. The results show that the proposed method can localize the needle accurately with a confidence of 99.2%.

Proceedings of the IEEE International Conference on Robotics and Automation (2019)

Towards robotic eye surgery: Marker-free, online hand-eye calibration using optical coherence tomography images

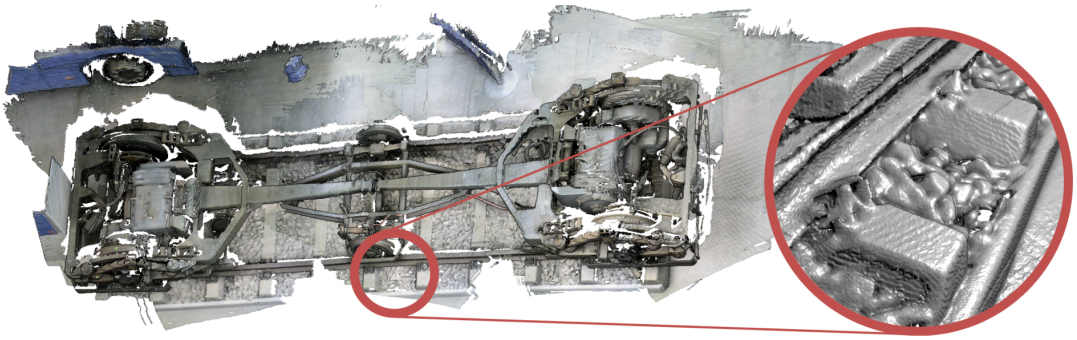
Mingchuan Zhou, Mahdi Hamad, Jakob Weiss, Abouzar Eslami, Kai Huang, Mathias Maier, Chris P. Lohmann, Nassir Navab, Alois Knoll, M. Ali Nasseri

Ophthalmic microsurgery is known to be a challenging operation, which requires very precise and dexterous manipulation. Image guided robot-assisted surgery (RAS) is a promising solution that brings significant improvements in outcomes and reduces the physical limitations of human surgeons. However, this technology must be further developed before it can be routinely used in clinics. One of the problems is the lack of proper calibration between the robotic manipulator and appropriate imaging device. In this work, we developed a flexible framework for hand-eye calibration of an ophthalmic robot with a microscope-integrated Optical Coherence Tomography (MIOCT) without any markers. The proposed method consists of three main steps: a) we estimate the OCT calibration parameters; b) with micro-scale displacements controlled by the robot, we detect and segment the needle tip in 3D-OCT volume; c) we find the transformation between the coordinate system of the OCT camera and the coordinate system of the robot. We verified the capability of our framework in ex-vivo pig eye experiments and compared the results with a reference method (marker-based). In all experiments, our method showed a small difference from the marker based method, with a mean calibration error of $9.2\,\mu\mathrm{m}$ and $7.0\,\mu\mathrm{m}$, respectively. Additionally, the noise test shows the robustness of the proposed method.

IEEE Robotics and Automation Letters Vol. 3 Issue 4 (2018)

Memory-Efficient Interactive Online Reconstruction From Depth Image Streams

Florian Reichl, Jakob Weiss, Rüdiger Westermann



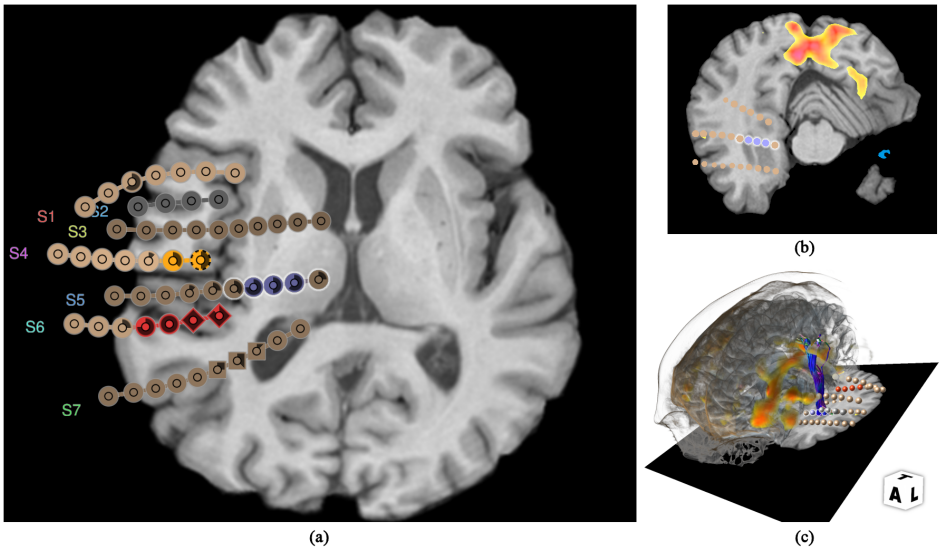
Online reconstruction of the underparts of a tramway. The scene is reconstructed at varying levels of resolutions of up to 1 mm (2.12 mm on average), requiring less than 10% of the memory that is required by alternative approaches.

We describe how the pipeline for 3D online reconstruction using commodity depth and image scanning hardware can be made scalable for large spatial extents and high scanning resolutions. Our modified pipeline requires less than 10% of the memory that is required by previous approaches at similar speed and resolution. To achieve this, we avoid storing a 3D distance field and weight map during online scene reconstruction. Instead, surface samples are binned into a high-resolution binary voxel grid. This grid is used in combination with caching and deferred processing of depth images to reconstruct the scene geometry. For pose estimation, GPU ray-casting is performed on the binary voxel grid. A one-to-one comparison to level-set ray-casting in a distance volume indicates slightly lower pose accuracy. To enable unlimited spatial extents and store acquired samples at the appropriate level of detail, we combine a hash map with a hierarchical tree representation.

Computer Graphics Forum Vol. 35 Issue 8 (2016)

Schematic Electrode Map for Navigation in Neuro Data Sets

Christian Schulte zu Berge, Jakob Weiss, Nassir Navab



(a) (b) (c)

(a) Our proposed schematic electrode map with the electrode glyph showing an overview over the depth electrode configuration. Furthermore, it serves as navigation tool for linked multi-modal visualizations, e.g. (b) definition of the 2D MPR plane; (b) camera placement or DTI fiber filtering in 3D volume rendering.

Neuro resection surgery is one of the last resorts when treating epilepsy patients where conservative treatment shows no effect on seizure reduction. However, due to the severity of the surgery, the resection planning has to be as precise as possible in order to avoid harming any critical anatomy. The tight time constraints in clinal routine demand for a highly optimized workflow. In this work, we therefore introduce a novel visualization in order to simplify the navigation in the complex multi-modal neuro data sets and support the clinician with the planning procedure. We propose a schematic electrode map based on a force-directed graph model providing an intuitive overview over the topology of the implanted depth electrode configuration. To further facilitate the planning workflow, our carefully designed electrode glyph supports different scalar, nominal and binary annotations augmenting the view with additional information. Brushing and linking techniques allow for easy mapping of the EEG data to the corresponding anatomy, as well as for straight-forward navigation within the visualization of the anatomical and functional imaging modalities in order to identify the origin and spread of the seizure. Our results show that the proposed graph layouting method successfully removes occlusions of the projected electrodes while maintaining the original topology of the depth electrode configuration. Initial discussions with clinicians and the application to clinical data further show the effectiveness of our methods.

Eurographics Workshop on Visual Computing for Biology and Medicine (2015)

An Open Source Multimodal Image-guided Prostate Biopsy Framework

Amit Shah, Oliver Zettinig, Tobias Maurer, Christina Precup, Christian Schulte zu Berge, Jakob Weiss, Benjamin Frisch, Nassir Navab

Although various modalities are used in prostate cancer imaging,transrectal ultrasound (TRUS) guided biopsy remains the gold standard for diagnosis. However, TRUS suffers from low sensitivity, leading to an elevated rate of false negative results. Magnetic Resonance Imaging (MRI) on the other hand provides currently the most accurate image based evaluation of the prostate. Thus, TRUS/MRI fusion image-guided biopsy has evolved to be the method of choice to circumvent the limitations of TRUS-only biopsy. Most commercial frameworks that offer such a solution rely on rigid TRUS/MRI fusion and rarely use additional information from other modalities such as Positron Emission Tomography (PET). Other frameworks require long interaction times and are complex to integrate with the clinical workflow. Available solutions are not fully able to meet the clinical requirements of speed and high precision at low cost simultaneously. We introduce an open source fusion biopsy framework that is low cost, simple to use and has minimal overhead in clinical workflow. Hence, it is ideal as a research platform for the implementation and rapid bench to bedside translation of new image registration and visualization approaches. We present the current status of the framework that uses pre-interventional PET and MRI rigidly registered with 3D TRUS for prostate biopsy guidance and discuss results from first clinical cases.

Workshop on Clinical Image-based Procedures: Translational Research in Medical Imaging (2014)

Bibliography

[1] D. C. Adler, T. H. Ko, and J. G. Fujimoto. “Speckle reduction in optical coherence tomography images by use of a spatially adaptive wavelet filter”. In: *Optics Letters* 29.24 (Dec. 2004), p. 2878 (cit. on p. 28).

[2] M. Allan, P.-L. Chang, S. Ourselin, et al. “Image based surgical instrument pose estimation with multi-class labelling and optical flow”. In: *MICCAI*. Springer. 2015, pp. 331–338 (cit. on p. 66).

[3] A. Apicella, F. Donnarumma, F. Isgrò, and R. Prevete. “A survey on modern trainable activation functions”. In: (May 2020). arXiv: 2005.00817 (cit. on p. 117).

[4] F. J. Ascaso, J. Lizana, and J. A. Cristóbal. “Cataract surgery in ancient Egypt”. In: *Journal of Cataract & Refractive Surgery* 35.3 (2009), pp. 607–608 (cit. on pp. 3, 5).

[5] H. Askari Poor. “Input device optimization to control an eye surgical robot for achieving an intuitive ophthalmic surgery”. Master’s Thesis. Politechnic Institute Milano, 2019 (cit. on p. 20).

[6] S. Asrani, L. Essaid, B. D. Alder, and C. Santiago-Turla. “Artifacts in spectral-domain optical coherence tomography measurements in glaucoma”. In: *JAMA Ophthalmology* 132.4 (2014), pp. 396–402 (cit. on p. 27).

[7] J. Aum, J.-h. Kim, and J. Jeong. “Effective speckle noise suppression in optical coherence tomography images using nonlocal means denoising filter with double Gaussian anisotropic kernels”. In: *Applied Optics* 54.13 (May 2015), p. D43 (cit. on p. 28).

[8] Y. M. Baek, S. Tanaka, H. Kanako, et al. “Full state visual forceps tracking under a microscope using projective contour models”. In: *Proc. of IEEE ICRA, pp. 2919–2925 (2012)* (cit. on p. 66).

[9] A. Barthel, D. Trematerra, M. A. Nasseri, et al. “Haptic interface for robot-assisted ophthalmic surgery”. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* 2015-November (2015), pp. 4906–4909 (cit. on p. 21).

[10] B. Baumann, C. W. Merkle, R. A. Leitgeb, et al. “Signal averaging improves signal-to-noise in OCT images: But which approach works best, and when?” In: *Biomedical Optics Express* 10.11 (Nov. 2019), p. 5755 (cit. on p. 98).

[11] H. Bay, T. Tuytelaars, and L. Van Gool. “Surf: Speeded up robust features”. In: *European conference on computer vision*. Springer. 2006, pp. 404–417 (cit. on p. 102).

[12] B. C. Becker, R. A. Maclachlan, L. A. Lobes, G. D. Hager, and C. N. Riviere. “Vision-based control of a handheld surgical micromanipulator with virtual fixtures”. In: *IEEE Transactions on Robotics* 29.3 (2013), pp. 674–683 (cit. on p. 19).

[13] J. M. Beer, A. D. Fisk, and W. A. Rogers. “Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction”. In: *Journal of Human-Robot Interaction* 3.2 (June 2014), p. 74 (cit. on p. 37).

[14] M. Ben Gayed, A. Guerrouad, C. Diaz, B. Lepers, and P. Vidal. “An advanced control micromanipulator for surgical applications”. In: *International system sciences* 13 (1987), pp. 123–34 (cit. on pp. 18, 20).

[15] L. Bennett. “Topical Versus Systemic Ocular Drug Delivery”. In: *Ocular Drug Delivery: Advances, Challenges and Applications*. Ed. by R. T. Addo. Cham: Springer International Publishing, 2016, pp. 53–74 (cit. on p. 16).

[16] M. Berger, J. Li, and J. A. Levine. “A Generative Model for Volume Rendering”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.4 (Apr. 2019), pp. 1636–1650. arXiv: 1710.09545 (cit. on p. 114).

[17] P. S. Bhende and A. Lobo. “Micro Incision Vitrectomy Surgery (MIVS): An Overview”. In: *Sci J Med & Vis Res Foun* XXXIII.2 (2015), pp. 57–60 (cit. on pp. 11, 15).

[18] I. D. Bleicher, M. Jackson-Atogi, C. Viehland, H. Gabr, J. A. Izatt, and C. A. Toth. “Depth-Based, Motion-Stabilized Colorization of Microscope-Integrated Optical Coherence Tomography Volumes for Microscope-Independent Microsurgery”. In: *Transl. Vis. Sci. Technol.* 7.6 (Nov. 2018), p. 1 (cit. on pp. 82, 84, 89, 90, 99).

[19] S. Borkovkina, A. Camino, W. Janpongsri, M. V. Sarunic, and Y. Jian. “Real-time retinal layer segmentation of OCT volumes with GPU accelerated inferencing using a compressed, low-latency neural network”. In: *Biomedical Optics Express* 11.7 (July 2020), p. 3968 (cit. on p. 92).

[20] A. Britten, P. Matten, M. Niederleithner, et al. “Versatile MEMS-VCSEL SS-OCT engine: full eye biometry to 4D imaging at MHz A-scan rates”. In: *Optical Coherence Tomography and Coherence Domain Optical Methods in Biomedicine XXV*. Ed. by J. A. Izatt and J. G. Fujimoto. SPIE, Mar. 2021, p. 16 (cit. on pp. 24, 91, 97).

[21] J. Brooke et al. “SUS-A quick and dirty usability scale”. In: *Usability evaluation in industry* 189.194 (1996), pp. 4–7 (cit. on p. 56).

[22] S. Bruckner and M. E. Gröller. “Instant Volume Visualization using Maximum Intensity Difference Accumulation”. In: *Computer Graphics Forum* 28.3 (2009), pp. 775–782 (cit. on pp. 81, 116).

[23] Carl Zeiss Meditec AG. *ZEISS OPMI LUMERA 700 (Marketing Material)*. Tech. rep. 2019 (cit. on p. 12).

[24] O. M. Carrasco-Zevallos, B. Keller, C. Viehland, et al. “Live volumetric (4D) visualization and guidance of in vivo human ophthalmic surgery with intraoperative optical coherence tomography”. In: *Scientific Reports* 6.1 (Oct. 2016), p. 31689 (cit. on pp. 26, 27, 81, 82).

[25] O. M. Carrasco-Zevallos, C. Viehland, B. Keller, R. P. McNabb, A. N. Kuo, and J. A. Izatt. “Constant linear velocity spiral scanning for near video rate 4D OCT ophthalmic and surgical imaging with isotropic transverse sampling”. In: *Biomedical Optics Express* 9.10 (2018), p. 5052 (cit. on pp. 24, 69, 92, 97).

[26] O. M. Carrasco-Zevallos, C. Viehland, B. Keller, et al. “Review of intraoperative optical coherence tomography: technology and applications [Invited].” In: *Biomed. Opt. Express* 8.3 (Mar. 2017), pp. 1607–1637 (cit. on pp. 22, 25).

[27] O. M. Carrasco-Zevallos, B. Keller, C. Viehland, et al. “Real-time 4D visualization of surgical maneuvers with 100kHz swept-source Microscope Integrated Optical Coherence Tomography (MIOCT) in model eyes”. In: *Investigative Ophthalmology & Visual Science* 55.13 (2014), p. 1633 (cit. on p. 26).

[28] K. V. Chalam and S. Gasparian. “Successful delivery of subretinal aflibercept (new surgical technique) for the treatment of submacular hemorrhage in idiopathic polypoidal choroidal vasculopathy”. In: *Journal of Surgical Case Reports* 2021.8 (Aug. 2021). rjab358. eprint: https://academic.oup.com/jscr/article-pdf/2021/8/rjab358/39763653/rjab358.pdf (cit. on p. 17).

[29] S. Charles. “Illumination and phototoxicity issues in vitreoretinal surgery”. In: *Retina* 28.1 (2008), pp. 1–4 (cit. on p. 14).

[30] S. Charles, J. Calzada, and B. Wood. *Vitreous Microsurgery*. Fifth Edition. Lippincott Williams & Wilkins, 2011, p. 251 (cit. on p. 15).

[31] G. Chatzipirpiridis, O. Ergeneman, J. Pokki, et al. “Electroforming of Implantable Tubular Magnetic Microrobots for Wireless Ophthalmologic Applications”. In: *Advanced Healthcare Materials* 4.2 (Jan. 2015), pp. 209–214 (cit. on p. 21).

[32] M. Chen, J. C. Gee, J. L. Prince, and G. K. Aguirre. “2D Modeling and Correction of Fan-Beam Scan Geometry in OCT”. In: *Computational Pathology and Ophthalmic Medical Image Analysis* (Sept. 2018), pp. 328–335 (cit. on p. 29).

[33] G. W. Cheon, Y. Huang, J. Cha, P. L. Gehlbach, and J. U. Kang. “Accurate real-time depth control for CP-SSOCT distal sensor based handheld microsurgery tools”. In: *Biomedical Optics Express* 6.5 (May 2015), p. 1942 (cit. on p. 24).

[34] Chia-Kai Liang, Li-Wen Chang, and H. Chen. “Analysis and Compensation of Rolling Shutter Effect”. In: *IEEE Transactions on Image Processing* 17.8 (Aug. 2008), pp. 1323–1330 (cit. on p. 32).

[35] C. Correa and Kwan-Liu Ma. “Size-based Transfer Functions: A New Volume Exploration Technique”. In: *IEEE Transactions on Visualization and Computer Graphics* 14.6 (Nov. 2008), pp. 1380–1387 (cit. on p. 112).

[36] C. D. Correa and K. L. Ma. “The occlusion spectrum for volume classification and visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 1465–1472 (cit. on p. 112).

[37] S. Cotin, J. Keppi, J. Allard, R. Bessard, H. Courtecuisse, and D. Gaucher. “Project RESET. REtinal Surgery systEm for Training”. In: *Acta Ophthalmologica* 93 (Oct. 2015), n/a–n/a (cit. on pp. 50, 53, 59).

[38] J. Danskin and P. Hanrahan. “Fast algorithms for volume ray tracing”. In: *Proceedings of the 1992 workshop on Volume visualization - VVS ’92*. 4. New York, New York, USA: ACM Press, 1992, pp. 91–98 (cit. on p. 114).

[39] F. de Moura Pinto and C. M. D. S. Freitas. “Importance-Aware Composition for Illustrative Volume Rendering”. In: *2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, Aug. 2010, pp. 134–141 (cit. on p. 116).

[40] F. De Moura Pinto and C. M. Freitas. “Design of Multi-dimensional Transfer Functions Using Dimensional Reduction”. In: *Eurographics/IEEE-VGTC Symposium on Visualization* February (2007). Ed. by K. Museth, T. Möller, and A. Ynnermann (cit. on pp. 113, 116).

[41] J. Díaz, T. Ropinski, I. Navazo, E. Gobbetti, and P. P. Vázquez. “An experimental study on the effects of shading in 3D perception of volumetric models”. In: *Visual Computer* 33.1 (2017), pp. 47–61 (cit. on pp. 54, 86).

[42] M. Draelos, B. Keller, C. Viehland, O. M. Carrasco-Zevallos, A. Kuo, and J. Izatt. “Real-time visualization and interaction with static and live optical coherence tomography volumes in immersive virtual reality”. In: *Biomed Opt Express* 9.6 (June 2018), p. 2825 (cit. on p. 27).

[43] M. Draelos, B. Keller, C. Viehland, O. M. Carrasco-Zevallos, A. Kuo, and J. Izatt. “Real-time visualization and interaction with static and live optical coherence tomography volumes in immersive virtual reality”. In: *Biomed. Opt. Express* 9.6 (June 2018), p. 2825 (cit. on pp. 51, 53).

[44] R. A. Drebin, L. Carpenter, and P. Hanrahan. “Volume rendering”. In: *ACM SIGGRAPH Computer Graphics* 22.4 (Aug. 1988), pp. 65–74 (cit. on pp. 81, 112).

[45] W. Drexler and J. Fujimoto. “Introduction to Optical Coherence Tomography”. In: *Optical Coherence Tomography*. 1. 2008, pp. 1–45 (cit. on pp. 21, 23, 24).

[46] W. Drexler and J. G. Fujimoto. *Optical Coherence Tomography*. Ed. by W. Drexler and J. G. Fujimoto. Biological and Medical Physics, Biomedical Engineering. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008 (cit. on p. 21).

[47] E. Dumortier, S. Cotin, J. Dequidt, and J.-F. Rouland. "A Prototype of Simulation System for Cataract Surgery Training". In: *European Society of Cataract & Refractive Surgeons Winter Meeting*. 2011 (cit. on p. 50).

[48] M. Dutra-Medeiros, J. Nascimento, J. Henriques, et al. "Three-Dimensional Head-Mounted Display System for Ophthalmic Surgical Procedures". In: *Retina* 37.7 (2017), pp. 1411–1414 (cit. on pp. 51, 53).

[49] T. L. Edwards, K. Xue, H. C. Meenink, et al. "First-in-human study of the safety and viability of intraocular robotic surgery". In: *Nature Biomedical Engineering* 2.9 (2018), pp. 649–656 (cit. on pp. 18, 20).

[50] J. Ehlers. "Intraoperative optical coherence tomography: past, present, and future". In: *Eye* 30.2 (2016), pp. 193–201 (cit. on pp. 4, 67).

[51] J. P. Ehlers, Y. K. Tao, S. Farsiu, R. Maldonado, J. A. Izatt, and C. A. Toth. "Integration of a spectral domain optical coherence tomography system into a surgical microscope for intraoperative imaging". In: *Investigative ophthalmology & visual science* 52.6 (2011), pp. 3153–3159 (cit. on p. 4).

[52] J. P. Ehlers, A. Uchida, and S. K. Srivastava. "Intraoperative optical coherence tomography-compatible surgical instruments for real-time image-guided ophthalmic surgery". In: *British Journal of Ophthalmology* 101.10 (2017), pp. 1306–1308 (cit. on p. 28).

[53] J. P. Ehlers, J. Goshe, W. J. Dupps, et al. "Determination of Feasibility and Utility of Microscope-Integrated Optical Coherence Tomography During Ophthalmic Surgery". In: *JAMA Ophthalmology* 133.10 (Oct. 2015), p. 1124 (cit. on p. 25).

[54] J. P. Ehlers, M. Khan, D. Petkovsek, et al. "Outcomes of Intraoperative OCT-Assisted Epiretinal Membrane Surgery from the PIONEER Study". In: *Ophthalmology Retina* (July 2017) (cit. on p. 79).

[55] J. P. Ehlers, Y. S. Modi, P. E. Pecen, et al. "The DISCOVER Study 3-Year Results". In: *Ophthalmology* 125.7 (July 2018), pp. 1014–1027 (cit. on pp. 22, 97).

[56] J. P. Ehlers, A. Uchida, and S. K. Srivastava. "Intraoperative optical coherence tomography-compatible surgical instruments for real-time image-guided ophthalmic surgery". In: *British Journal of Ophthalmology* 101.10 (2017), pp. 1306–1308 (cit. on p. 99).

[57] J. P. Ehlers, A. Uchida, and S. K. Srivastava. "THE INTEGRATIVE SURGICAL THEATER: Combining Intraoperative Optical Coherence Tomography and 3D Digital Visualization for Vitreoretinal Surgery in the DISCOVER Study". In: *Retina* 38 (Sept. 2018), S88–S96 (cit. on pp. 4, 81).

[58] D. Engel and T. Ropinski. "Deep Volumetric Ambient Occlusion". In: (Aug. 2020). arXiv: 2008.08345 (cit. on pp. 113, 130).

[59] K. Engel, M. Kraus, and T. Ertl. "High-quality pre-integrated volume rendering using hardware-accelerated pixel shading". In: *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS workshop on Graphics hardware - HWWS '01*. New York, New York, USA: ACM Press, 2001, pp. 9–16 (cit. on p. 120).

[60] O. Ergeneman, C. Bergeles, M. P. Kummer, J. J. Abbott, and B. J. Nelson. "Wireless Intraocular Microrobots: Opportunities and Challenges". In: *Surgical Robotics: Systems Applications and Visions*. Ed. by J. Rosen, B. Hannaford, and R. M. Satava. 2011, pp. 1–819 (cit. on p. 21).

[61] E. Ezra. "Idiopathic full thickness macular hole: Natural history and pathogenesis". In: vol. 85. 1. BMJ Publishing Group Ltd, Jan. 2001, pp. 102–108 (cit. on p. 15).

[62] K. C. Fan and A. M. Berrocal. “Surgical Techniques for Retinal Gene Therapy Delivery”. In: *Retinal Physician* 17 (Feb. 2020), pp. 26–28 (cit. on p. 6).

[63] Fang-Yu Lin, C. Bergeles, and Guang-Zhong Yang. “Biometry-based concentric tubes robot for vitreoretinal surgery”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Aug. 2015, pp. 5280–5284 (cit. on p. 20).

[64] A. F. Fercher, C. K. Hitzenberger, W. Drexler, G. Kamp, and H. Sattmann. “In vivo optical coherence tomography”. In: *American journal of ophthalmology* (1993) (cit. on p. 21).

[65] C. J. Flaxel, R. A. Adelman, S. T. Bailey, et al. “Idiopathic Macular Hole Preferred Practice Pattern”. In: *Ophthalmology* 127.2 (Feb. 2020), P184–P222 (cit. on p. 16).

[66] A. B. Fuente. “Retinal pjototoxicity after macular hole surgery induced by xenon light: A case series”. In: *Vision Pan-America, The Pan-American Journal of Ophthalmology* 12.1 (2013), pp. 17–20 (cit. on p. 14).

[67] Y. Gan, W. Yao, K. M. Myers, and C. P. Hendon. “An automated 3D registration method for optical coherence tomography volumes”. In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Aug. 2014, pp. 3873–3876 (cit. on pp. 98, 100).

[68] L. Garcia Peraza Herrera, W. Li, C. Gruijthuijsen, et al. “Real-Time Segmentation of Non-Rigid Surgical Tools based on Deep Learning and Tracking”. In: *CARE workshop at International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016 (cit. on p. 66).

[69] A. Gijbels, K. Willekens, L. Esteveny, P. Stalmans, D. Reynaerts, and E. B. Vander Poorten. “Towards a clinically applicable robotic assistance system for retinal vein cannulation”. In: *Proceedings of the IEEE RAS and EMBS International Conference on Biomedical Robotics and Biomechatronics* 2016-July (2016), pp. 284–291 (cit. on p. 20).

[70] A. Gijbels, N. Wouters, P. Stalmans, H. Van Brussel, D. Reynaerts, and E. V. Poorten. “Design and realisation of a novel robotic manipulator for retinal surgery”. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Nov. 2013, pp. 3598–3603 (cit. on p. 19).

[71] L. Ginner, C. Blatter, D. Fechtig, T. Schmoll, M. Gröschl, and R. Leitgeb. “Wide-Field OCT Angiography at 400 KHz Utilizing Spectral Splitting”. In: *Photonics* 1.4 (2014), pp. 369–379 (cit. on pp. 91, 97).

[72] K. Grace, J. Colgate, M. Glucksberg, and J. Chun. “A six degree of freedom micromanipulator for ophthalmic surgery”. In: *Proceedings IEEE International Conference on Robotics and Automation (1993)*. IEEE Comput. Soc. Press, 1993, pp. 630–635 (cit. on p. 20).

[73] I. Grulkowski, J. J. Liu, B. Potsaid, V. Jayaraman, A. E. Cable, and J. G. Fujimoto. “Ultrahigh Speed OCT”. In: *Optical Coherence Tomography*. Ed. by W. Drexler and J. G. Fujimoto. Cham: Springer International Publishing, 2015, pp. 319–356 (cit. on p. 24).

[74] I. Grulkowski, J. J. Liu, B. Potsaid, et al. “Retinal, anterior segment and full eye imaging using ultrahigh speed swept source OCT with vertical-cavity surface emitting lasers”. In: *Biomedical Optics Express* 3.11 (Nov. 2012), p. 2733 (cit. on p. 97).

[75] A. Guerrouad and P. Vidal. “SMOS: stereotaxical microtelemanipulator for ocular surgery”. In: *Images of the Twenty-First Century. Proceedings of the Annual International Engineering in Medicine and Biology Society*. IEEE, 1989, pp. 879–880 (cit. on pp. 18, 20).

[76] P. K. Gupta, P. S. Jensen, and E. De Juan. “Surgical forces and tactile perception during retinal microsurgery”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 1679. 1999, pp. 1218–1225 (cit. on p. 18).

[77] M. Hadwiger, A. Kratz, C. Sigg, and K. Bühler. “GPU-Accelerated Deep Shadow Maps for Direct Volume Rendering”. In: *Graphics Hardware* (2006) (cit. on p. 32).

[78] M. Hadwiger, P. Ljung, C. R. Salama, and T. Ropinski. “Advanced illumination techniques for GPU volume raycasting”. In: *ACM SIGGRAPH ASIA 2008 Courses* (2008), pp. 1–166 (cit. on p. 114).

[79] M. Häggström. *Fundus photograph of normal right eye*. CC0. 2014 (cit. on p. 10).

[80] I. C. Han and G. J. Jaffe. “Evaluation of Artifacts Associated with Macular Spectral-Domain Optical Coherence Tomography”. In: *Ophthalmology* 117.6 (June 2010), 1177–1189.e4 (cit. on p. 27).

[81] Hanqi Guo, Ningyu Mao, and Xiaoru Yuan. “WYSIWYG (What You See is What You Get) Volume Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (Dec. 2011), pp. 2106–2114 (cit. on pp. 113, 125).

[82] R. C. Harwell and R. L. Ferguson. “Physiologic tremor and microsurgery”. In: *Microsurgery* 4.3 (1983), pp. 187–192 (cit. on p. 18).

[83] C.-Y. He, L. Huang, Y. Yang, Q.-F. Liang, and Y.-K. Li. “Research and Realization of a Master-Slave Robotic System for Retinal Vascular Bypass Surgery”. In: *Chinese Journal of Mechanical Engineering* 31.1 (Dec. 2018), p. 78 (cit. on p. 20).

[84] K. He, X. Zhang, S. Ren, and J. Sun. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV] (cit. on p. 20).

[85] W. He, J. Wang, H. Guo, et al. “InSituNet: Deep Image Synthesis for Parameter Space Exploration of Ensemble Simulations”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 23–33. arXiv: 1908.00407 (cit. on p. 113).

[86] X. He, M. A. Balicki, J. U. Kang, et al. “Force sensing micro-forceps with integrated fiber Bragg grating for vitreoretinal surgery”. In: *Optical Fibers and Sensors for Medical Diagnostics and Treatment Applications XII* 8218 (2012), 82180W (cit. on p. 20).

[87] X. He, D. Roppenecker, D. Gierlach, et al. “Toward Clinically Applicable Steady-Hand Eye Robot for Vitreoretinal Surgery”. In: *Volume 2: Biomedical and Biotechnology*. American Society of Mechanical Engineers, Nov. 2012, pp. 145–153 (cit. on p. 20).

[88] HelpMeSee Inc. *Eye Surgery Simulator*. 2021 (cit. on p. 50).

[89] F. Hernell, P. Ljung, and A. Ynnerman. “Local ambient occlusion in direct volume rendering”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.4 (2010), pp. 548–559 (cit. on p. 32).

[90] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. “GANs trained by a two time-scale update rule converge to a local Nash equilibrium”. In: *Advances in Neural Information Processing Systems*. Vol. 2017-Decem. 2017, pp. 6627–6638 (cit. on p. 122).

[91] J. Ho, D. P. E. Castro, L. C. Castro, et al. “Clinical Assessment of Mirror Artifacts in Spectral-Domain Optical Coherence Tomography”. In: *Investigative Opthalmology & Visual Science* 51.7 (July 2010), p. 3714 (cit. on p. 28).

[92] F. Hong, C. Liu, and X. Yuan. “DNN-VolVis: Interactive volume visualization supported by deep neural network”. In: *IEEE Pacific Visualization Symposium* 2019-April (2019), pp. 282–291 (cit. on p. 113).

[93] D. Huang, E. Swanson, C. Lin, et al. “Optical coherence tomography”. In: *Science* 254.5035 (Nov. 1991), pp. 1178–1181 (cit. on p. 21).

[94] I. W. Hunter, T. D. Doukoglou, S. R. Lafontaine, et al. “A Teleoperated Microsurgical Robot and Associated Virtual Environment for Eye Surgery”. In: *Presence: Teleoperators and Virtual Environments* 2.4 (1993), pp. 265–280 (cit. on pp. 20, 50).

[95] Y. Ida, N. Sugita, T. Ueta, Y. Tamaki, K. Tanimoto, and M. Mitsuishi. “Microsurgical robotic system for vitreoretinal surgery”. In: *International Journal of Computer Assisted Radiology and Surgery* 7.1 (Jan. 2012), pp. 27–34 (cit. on p. 20).

[96] J. A. Izatt and M. A. Choma. “Theory of Optical Coherence Tomography”. In: *Optical Coherence Tomogrpahy*. Ed. by W. Drexler and J. G. Fujimoto. 2008, pp. 47–72 (cit. on p. 23).

[97] S. Jain, W. Griffin, A. Godil, J. W. Bullard, J. Terrill, and A. Varshney. “Compressed Volume Rendering using Deep Learning”. In: *Proceedings of the Large Scale Data Analysis and Visualization (LDAV) Symposium*. Phoenix, AZ, Oct. 2017 (cit. on p. 113).

[98] A. H. Jazwinski. *Stochastic processes and filtering theory*. Courier Corporation, 2007 (cit. on p. 71).

[99] Jmarchn. *Diagram of the human eye in English. It shows the lower part of the right eye after a central and horizontal section.* CC BY-SA 3.0. 2007 (cit. on p. 10).

[100] J. Johnson, A. Alahi, and L. Fei-Fei. “Perceptual losses for real-time style transfer and super-resolution”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9906 LNCS. 2016, pp. 694–711. arXiv: 1603.08155v1 (cit. on p. 133).

[101] D. Jönsson, J. Kronander, T. Ropinski, and A. Ynnerman. “Historygrams: Enabling Interactive Global Illumination in Direct Volume Rendering using Photon Mapping”. In: *IEEE Trans Vis Comput Graph* 18.12 (2012), pp. 2364–2371 (cit. on p. 32).

[102] D. Jönsson, E. Sundén, A. Ynnerman, and T. Ropinski. “A Survey of Volumetric Illumination Techniques for Interactive Volume Rendering”. In: *Computer Graphics Forum* 33.1 (2014), pp. 27–51 (cit. on p. 32).

[103] J. U. Kang and G. W. Cheon. “Demonstration of subretinal injection using common-path swept source OCT guided microinjector”. In: *Applied Sciences (Switzerland)* 8.8 (2018) (cit. on p. 19).

[104] H. Kato, D. Beker, M. Morariu, et al. “Differentiable Rendering: A Survey”. In: (June 2020). arXiv: 2006.12057 (cit. on p. 113).

[105] Y. M. Khalifa, D. Bogorad, V. Gibson, J. Peifer, and J. Nussbaum. “Virtual Reality in Ophthalmology Training”. In: *Survey of Ophthalmology* 51.3 (2006), pp. 259–273 (cit. on p. 50).

[106] G. Kindlmann, R. Whitaker, T. Tasdizen, and T. Moller. “Curvature-based transfer functions for direct volume rendering: methods and applications”. In: (2003), pp. 513–520 (cit. on p. 112).

[107] D. P. Kingma and J. L. Ba. “Adam: A method for stochastic optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. 2015. arXiv: 1412.6980 (cit. on pp. 124, 128).

[108] M. Y. Kirillin, G. Farhat, E. A. Sergeeva, M. C. Kolios, and A. Vitkin. “Speckle statistics in OCT images: Monte Carlo simulations and experimental studies”. In: *Optics Letters* 39.12 (2014), p. 3472 (cit. on p. 27).

[109] J. Kniss, G. Kindlmann, and C. Hansen. “Interactive volume rendering using multi-dimensional transfer functions and direct manipulation widgets”. In: *Proceedings Visualization, 2001. VIS ’01.* IEEE, 2001, pp. 255–562 (cit. on p. 112).

[110] J. Kniss, S. Premože, M. Ikits, A. Lefohn, C. Hansen, and E. Praun. “Gaussian Transfer Functions for Multi-Field Volume Visualization”. In: *Proceedings of the IEEE Visualization Conference* Vi (2003), pp. 497–504 (cit. on p. 116).

[111] J. P. Kolb, W. Draxinger, J. Klee, et al. “Live video rate volumetric OCT imaging of the retina with multi-MHz A-scan rates”. In: *PLOS ONE* 14.3 (Mar. 2019). Ed. by B. V. Bui, e0213144 (cit. on pp. 4, 24, 34, 82, 92, 97, 98, 106).

[112] M. F. Kraus, B. Potsaid, M. A. Mayer, et al. “Motion correction in optical coherence tomography volumes on a per A-scan basis using orthogonal scan patterns.” In: *Biomedical optics express* 3.6 (June 2012), pp. 1182–99 (cit. on p. 98).

[113] T. Kroes, F. H. Post, C. P. Botha, S. Longworth, and M. Yu. “Exposure Render: An Interactive Photo-Realistic Volume Rendering Framework”. In: *PLoS ONE* 7.7 (July 2012). Ed. by X.-N. Zuo, e38586 (cit. on pp. 32, 33, 82).

[114] J. Krüger and R. Westermann. “Acceleration techniques for GPU-based volume rendering”. In: *Proceedings of the 14th IEEE Visualization 2003 (VIS'03)* (2004), pp. 287–292 (cit. on pp. 86, 93, 112).

[115] J. Kugelman, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins. “Automatic segmentation of OCT retinal boundaries using recurrent neural networks and graph search”. In: *Biomedical optics express* 9.11 (2018), pp. 5759–5777 (cit. on p. 89).

[116] M. P. Kummer, J. J. Abbott, B. E. Kratochvil, R. Borer, A. Sengul, and B. J. Nelson. “OctoMag: An Electromagnetic System for 5-DOF Wireless Micromanipulation”. In: *IEEE Transactions on Robotics* 26.6 (Dec. 2010), pp. 1006–1017 (cit. on p. 21).

[117] T. Kurmann, P. M. Neila, X. Du, et al. “Simultaneous Recognition and Pose Estimation of Instruments in Minimally Invasive Surgery”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 505–513 (cit. on p. 66).

[118] I. Laina, N. Rieke, C. Rupprecht, et al. “Concurrent Segmentation and Localization for Tracking of Surgical Instruments”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention 2017*. The first two authors contributed equally to this paper. Springer. 2017, pp. 664–672 (cit. on p. 66).

[119] T. Leng, J. M. Miller, K. V. Bilbao, D. V. Palanker, P. Huie, and M. S. Blumenkranz. “The Chick Chorioallantoic Membrane as a Model Tissue for Surgical Retinal Research and Simulation”. In: *Retina* 24.3 (June 2004), pp. 427–434 (cit. on p. 14).

[120] M. Levoy. “Display of surfaces from volume data”. In: *IEEE Computer Graphics and Applications* 8.3 (May 1988), pp. 29–37 (cit. on p. 31).

[121] J. Q. Li, T. Welchowski, M. Schmid, M. M. Mauschitz, F. G. Holz, and R. P. Finger. “Prevalence and incidence of age-related macular degeneration in Europe: a systematic review and meta-analysis”. In: *British Journal of Ophthalmology* 104.8 (2020), pp. 1077–1084 (cit. on p. 6).

[122] M. Li, R. Idoughi, B. Choudhury, and W. Heidrich. “Statistical model for OCT image denoising”. In: *Biomedical Optics Express* 8.9 (Sept. 2017), p. 3903 (cit. on p. 28).

[123] T. M. Li, M. Aittala, F. Durand, and J. Lehtinen. “Differentiable Monte Carlo ray tracing through edge sampling”. In: *SIGGRAPH Asia 2018 Technical Papers, SIGGRAPH Asia 2018*. 2018 (cit. on p. 113).

[124] X. Li, L. Wei, X. Dong, et al. “Microscope-integrated optical coherence tomography for image-aided positioning of glaucoma surgery”. In: *Journal of Biomedical Optics* 20.07 (July 2015), p. 1 (cit. on p. 26).

[125] O. Liba, M. D. Lew, E. D. SoRelle, et al. “Speckle-modulating optical coherence tomography in living mice and humans”. In: *Nature Communications* 8.1 (Aug. 2017), p. 15845 (cit. on p. 28).

[126] S. Liu, Y. Zhang, S. Peng, B. Shi, M. Pollefeys, and Z. Cui. “DIST: Rendering Deep Implicit Signed Distance Function With Differentiable Sphere Tracing”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020, pp. 2016–2025. arXiv: 1911.13225 (cit. on p. 114).

[127] S. Liu, W. Chen, T. Li, and H. Li. “Soft rasterizer: A differentiable renderer for image-based 3D reasoning”. In: *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 2019-Octob. 2019, pp. 7707–7716. arXiv: 1904.01786 (cit. on p. 113).

[128] Y. Liu, X. Xu, and F. Li. “Image Feature Matching Based on Deep Learning”. In: *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*. IEEE, Dec. 2018, pp. 1752–1756 (cit. on p. 107).

[129] P. Ljung, J. Krüger, E. Groller, M. Hadwiger, C. D. Hansen, and A. Ynnerman. “State of the Art in Transfer Functions for Direct Volume Rendering”. In: *Computer Graphics Forum* 35.3 (2016), pp. 669–691 (cit. on pp. 31, 110, 112).

[130] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh. “Neural volumes: Learning dynamic renderable volumes from images”. In: *ACM Transactions on Graphics* 38.4 (2019), p. 14. arXiv: 1906.07751 (cit. on p. 114).

[131] I. Loshchilov and F. Hutter. “Decoupled Weight Decay Regularization”. In: *International Conference on Learning Representations*. 2019 (cit. on p. 106).

[132] G. Loubet, N. Holzschuch, and W. Jakob. “Reparame-terizing Discontinuous Integrands for Differentiable Rendering”. In: *ACM Trans. Graph* 38 (2019), p. 14 (cit. on p. 113).

[133] A. Mablekos-Alexiou, S. Ourselin, L. Da Cruz, and C. Bergeles. “Requirements Based Design and End-to-End Dynamic Modeling of a Robotic Tool for Vitreoretinal Surgery”. In: *Proceedings - IEEE International Conference on Robotics and Automation* (2018), pp. 135–141 (cit. on p. 20).

[134] R. Machemer, H. Buettner, E. W. Norton, and J. M. Parel. “Vitrectomy: a pars plana approach”. In: *Trams Am Acad Ophthalmol Otolaryngol* 75.4 (1971), pp. 813–820 (cit. on p. 3).

[135] R. A. MacLachlan, B. C. Becker, J. C. Tabarés, G. W. Podnar, L. A. Lobes, and C. N. Riviere. “Micron: An actively stabilized handheld tool for microsurgery”. In: *IEEE Transactions on Robotics*. Vol. 28. 1. NIH Public Access, Feb. 2012, pp. 195–212 (cit. on p. 19).

[136] R. E. MacLaren and R. A. Pearson. “Stem cell therapy and the retina”. In: *Eye* 21.10 (2007), pp. 1352–1359 (cit. on pp. 16, 17).

[137] A. W. Mahoney, N. D. Nelson, E. M. Parsons, and J. J. Abbott. “Non-ideal behaviors of magnetically driven screws in soft tissue”. In: *IEEE International Conference on Intelligent Robots and Systems* (2012), pp. 3559–3564 (cit. on p. 21).

[138] P. M. Maloca, J. E. R. de Carvalho, T. Heeren, et al. “High-Performance Virtual Reality Volume Rendering of Original Optical Coherence Tomography Point-Cloud Data Enhanced With Real-Time Ray Casting.” In: *Translational vision science & technology* 7.4 (July 2018), p. 2 (cit. on pp. 50, 51).

[139] N. Max. “Optical models for direct volume rendering”. In: *IEEE Transactions on Visualization and Computer Graphics* 1.2 (June 1995), pp. 99–108 (cit. on pp. 30, 110, 114).

[140] N. Max. “Optical models for direct volume rendering”. In: *IEEE Transactions on Visualization and Computer Graphics* 1.2 (June 1995), pp. 99–108 (cit. on p. 30).

[141] Y. Mboussou. *Histology of Macula in OCT*. CC BY-SA 4.0. 2017 (cit. on p. 11).

[142] H. Meenink, R. Hendrix, G. Naus, et al. “Robot-assisted vitreoretinal surgery”. In: *Medical Robotics* (2012), pp. 185–209 (cit. on p. 20).

[143] S. Mehta and G. B. Hubbard. “Avoiding Neck Strain in Vitreoretinal Surgery”. In: *Retina* 33.2 (Feb. 2013), pp. 439–441 (cit. on p. 14).

[144] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”. In: (Mar. 2020). arXiv: 2003.08934 (cit. on p. 114).

[145] P. Milgram and F. Kishino. “A Taxonomy of Mixed Reality Visual Displays”. In: *IEICE Transactions on Information Systems* E77-D.12 (1994) (cit. on p. 52).

[146] F. Milletari, N. Navab, and S.-A. Ahmadi. “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation”. In: *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, Oct. 2016, pp. 565–571. arXiv: 1606.04797v1 (cit. on p. 118).

[147] P. Mitchell, W. Smith, T. Chey, Jie Jin Wang, and A. Chang. “Prevalence and associations of epiretinal membranes: The blue mountains eye study, Australia”. In: *Ophthalmology* 104.6 (1997), pp. 1033–1040 (cit. on p. 15).

[148] N. Moura-Coelho, J. Henriques, J. Nascimento, and M. Dutra-Medeiros. “Three-dimensional Display Systems in Ophthalmic Surgery – A Review”. In: *European Ophthalmic Review* 13.1 (2019) (cit. on pp. 14, 51).

[149] L. Mroz, H. Hauser, and E. Gröller. “Interactive High-Quality Maximum Intensity Projection”. In: *Computer Graphics Forum* 19.3 (Sept. 2000), pp. 341–350 (cit. on p. 81).

[150] M. Muja and D. G. Lowe. “Scalable nearest neighbor algorithms for high dimensional data”. In: *IEEE transactions on pattern analysis and machine intelligence* 36.11 (2014), pp. 2227–2240 (cit. on p. 102).

[151] M. Nagpal, A. Verma, and S. Goswami. “Micro-incision Vitrectomy Surgery – Past, Present and Future”. In: *European Ophthalmic Review* 09.01 (2015), p. 64 (cit. on pp. 11, 14).

[152] M. Nambi, P. S. Bernstein, and J. J. Abbott. “A Compact Telemanipulated Retinal-Surgery System that Uses Commercially Available Instruments with a Quick-Change Adapter”. In: *Journal of Medical Robotics Research* 01.02 (June 2016), p. 1630001 (cit. on p. 19).

[153] M. A. Nasseri, M. Eder, S. Nair, et al. “The introduction of a new robot for assistance in ophthalmic surgery”. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (2013), pp. 5682–5685 (cit. on pp. 21, 38).

[154] T. Nguyen-Phuoc, C. Li, Y. L. Yang, and S. Balaban. “Rendernet: A deep convolutional network for differentiable rendering from 3D shapes”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 7891–7901. arXiv: 1806.06575 (cit. on pp. 113, 118).

[155] M. Niemeijer, M. Garvin, K. Lee, B. Ginneken, M. Abramoff, and M. Sonka. “Registration of 3D spectral OCT volumes using 3D SIFT feature point matching”. In: *Proceedings of SPIE - The International Society for Optical Engineering* (Feb. 2009) (cit. on p. 98).

[156] M. Niemeijer, K. Lee, M. K. Garvin, M. D. Abràmoff, and M. Sonka. “Registration of 3D spectral OCT volumes combining ICP with a graph-based approach”. In: *Medical Imaging 2012: Image Processing*. Ed. by D. R. Haynor and S. Ourselin. Vol. 8314. International Society for Optics and Photonics. SPIE, 2012, pp. 378–386 (cit. on pp. 98, 100).

[157] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. “Differentiable Volumetric Rendering: Learning Implicit 3D Representations Without 3D Supervision”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020, pp. 3501–3512. arXiv: 1912.07372 (cit. on p. 114).

[158] M. Nimier-David, É. Polytechnique, F. De Lausanne, D. Vicini, T. Zeltner, and W. Jakob. “Mitsuba 2: A Retargetable Forward and Inverse Renderer”. In: *ACM Trans. Graph* 38.6 (2019), p. 17 (cit. on p. 113).

[159] G. Ometto, I. Moghul, G. Montesano, et al. “ReLayer: a Free, Online Tool for Extracting Retinal Thickness From Cross-Platform OCT Images”. In: *Translational Vision Science & Technology* 8.3 (May 2019), p. 25 (cit. on p. 89).

[160] L. Pan, L. Guan, and X. Chen. “Segmentation Guided Registration for 3D Spectral-Domain Optical Coherence Tomography Images”. In: *IEEE Access* 7 (2019), pp. 138833–138845 (cit. on pp. 98, 100, 103).

[161] Y. Peng, L. Tang, and Y. Zhou. “Subretinal Injection: A Review on the Novel Route of Therapeutic Delivery for Vitreoretinal Diseases”. In: *Ophthalmic Research* 58.4 (2017), pp. 217–226 (cit. on pp. 6, 16).

[162] M. Pharr, W. Jakob, and G. Humphreys. “Volume Scattering”. In: *Physically Based Rendering*. Elsevier, 2017, pp. 671–704 (cit. on p. 32).

[163] A. Podoleanu, I. Charalambous, L. Plesea, A. Dogariu, and R. Rosen. “Correction of distortions in optical coherence tomography imaging of the eye”. In: *Physics in Medicine and Biology* 49.7 (Apr. 2004), pp. 1277–1294 (cit. on p. 29).

[164] J. S. Pollack and N. Sabherwal. “Small gauge vitrectomy: operative techniques”. In: *Current opinion in ophthalmology* 30.3 (2019), pp. 159–164 (cit. on pp. 12, 14).

[165] B. Potsaid, I. Gorczynska, V. J. Srinivasan, et al. “Ultrahigh speed Spectral / Fourier domain OCT ophthalmic imaging at 70,000 to 312,500 axial scans per second”. In: *Optics Express* 16.19 (Sept. 2008), p. 15149 (cit. on p. 23).

[166] J. Probst, D. Hillmann, E. Lankenau, et al. “Optical coherence tomography with online visualization of more than seven rendered volumes per second”. In: *Journal of Biomedical Optics* 15.2 (2010), p. 026014 (cit. on p. 26).

[167] K. Rematas and V. Ferrari. “Neural Voxel Renderer: Learning an Accurate and Controllable Rendering Tool”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020, pp. 5416–5426. arXiv: 1912.04591 (cit. on p. 113).

[168] C. Rezk-Salama, M. Keller, and P. Kohlmann. “High-Level User Interfaces for Transfer Function Design with Semantics”. In: *IEEE Transactions on Visualization and Computer Graphics* 12.5 (Sept. 2006), pp. 1021–1028 (cit. on pp. 112, 116).

[169] R. Richa, M. Balicki, E. Meisner, R. Sznitman, R. Taylor, and G. Hager. “Visual tracking of surgical tools for proximity detection in retinal surgery”. In: *IPCAI, pp. 55–66 (2011)* (cit. on p. 66).

[170] N. Rieke, D. J. Tan, C. Amat di San Filippo, et al. “Real-time Localization of Articulated Surgical Instruments in Retinal Microsurgery”. In: *Medical Image Analysis* 34 (2016) (cit. on p. 66).

[171] N. Rieke, D. J. Tan, F. Tombari, et al. “Real-Time Online Adaption for Robust Instrument Tracking and Pose Estimation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 422–430 (cit. on p. 66).

[172] M. Roizenblatt, T. Edwards, and P. L. Gehlbach. “Robot-assisted vitreoretinal surgery: current perspectives”. In: *Robotic Surgery: Research and Reviews* Volume 5 (Feb. 2018), pp. 1–11 (cit. on p. 79).

[173] H. Roodaki, K. Filippatos, A. Eslami, and N. Navab. “Introducing Augmented Reality to Optical Coherence Tomography in Ophthalmic Microsurgery”. In: *2015 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, Sept. 2015, pp. 1–6 (cit. on pp. 50, 53).

[174] T. Ropinski, C. Doring, and C. Rezk-Salama. “Interactive volumetric lighting simulating scattering and shadowing”. In: (2010), pp. 169–176 (cit. on p. 32).

[175] T. Ropinski, J. Kasten, and K. H. Hinrichs. “Efficient shadows for GPU-based volume raycasting”. In: *WSCG 2008* (2008), pp. 17–24 (cit. on pp. 86, 113, 125, 128).

[176] T. R. Rosenblatt, D. Vail, N. Saroj, N. Boucher, D. M. Moshfeghi, and A. A. Moshfeghi. “Increasing Incidence and Prevalence of Common Retinal Diseases in Retina Practices Across the United States”. In: *Ophthalmic Surgery, Lasers and Imaging Retina* 52.1 (Jan. 2021), pp. 29–36 (cit. on p. 6).

[177] S. Röttger. *The Volume Library*. 2020 (cit. on p. 122).

[178] A. G. Roy, S. Conjeti, S. P. K. Karri, et al. “ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks”. In: *Biomedical Optics Express* 8.8 (Aug. 2017), p. 3627. eprint: 1704.02161 (cit. on p. 89).

[179] M. Ruiz, A. Bardera, I. Boada, I. Viola, M. Feixas, and M. Sbert. “Automatic transfer functions based on informational divergence”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 1932–1941 (cit. on p. 113).

[180] R. B. Rusu, N. Blodow, and M. Beetz. “Fast point feature histograms (FPFH) for 3D registration”. In: *2009 IEEE international conference on robotics and automation*. IEEE. 2009, pp. 3212–3217 (cit. on p. 102).

[181] Y. Sato, N. Shiraga, S. Nakajima, S. Tamura, and R. Kikinis. “Local maximum intensity projection (LMIP): A new rendering method for vascular visualization”. In: *J. Comput. Assist. Tomo.* 22.6 (1998), pp. 912–917 (cit. on p. 81).

[182] F. Schirrmacher, T. Köhler, J. Endres, et al. “Temporal and volumetric denoising via quantile sparse image prior”. In: *Medical Image Analysis* 48 (Aug. 2018), pp. 131–146 (cit. on p. 28).

[183] J. M. Schmitt, S. H. Xiang, and K. M. Yung. “Speckle in Optical Coherence Tomography”. In: *Journal of Biomedical Optics* 4.1 (1999), p. 95 (cit. on p. 27).

[184] C. Schulte zu Berge, M. Baust, A. Kapoor, and N. Navab. “Predicate-Based Focus-and-Context Visualization for 3D Ultrasound”. In: *IEEE Trans Vis Comput Graph* 20.12 (2014), pp. 2379–2387 (cit. on p. 113).

[185] C. Schulte zu Berge, A. Grunau, H. Mahmud, and N. Navab. “CAMPVis - A Game Engine-inspired Research Framework for Medical Imaging and Visualization”. Munich, 2014 (cit. on pp. 82, 106).

[186] X. Shen, F. Darmon, A. A. Efros, and M. Aubry. “RANSAC-Flow: Generic Two-Stage Image Alignment”. In: 2020, pp. 618–637 (cit. on p. 107).

[187] M. S. Singh, S. S. Park, T. A. Albini, et al. “Retinal stem cell transplantation: Balancing safety and potential”. In: *Progress in Retinal and Eye Research* 75.October 2018 (2020), p. 100779 (cit. on pp. 16, 17).

[188] V. Sitzmann, M. Zollhöfer, and G. Wetzstein. “Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations”. In: *Conference on Neural Information Processing Systems* (2019), pp. 1121–1132 (cit. on p. 114).

[189] R. Skarbez, M. Smith, and M. C. Whitton. “Revisiting Milgram and Kishino’s Reality-Virtuality Continuum”. In: *Frontiers in Virtual Reality* 2 (Mar. 2021) (cit. on pp. 52, 53, 59).

[190] T. Sohail, M. Pajaujis, S. E. Crawford, J. W. Chan, and T. Eke. “Face-to-face upright seated positioning for cataract surgery in patients unable to lie flat: Case series of 240 consecutive phacoemulsifications”. In: *Journal of Cataract & Refractive Surgery* 44.9 (2018), pp. 1116–1122 (cit. on p. 11).

[191] M. Sommersperger. “Real-Time Volumetric Registration for intraoperative OCT”. Technical University of Munich, 2018 (cit. on p. 102).

[192] M. Sommersperger, J. Weiss, M. Nasseri, P. Gehlbach, I. Iordachita, and N. Navab. “Real-time tool to layer distance estimation for robotic subretinal injection using intraoperative 4D OCT”. In: *Biomedical Optics Express* 12.2 (Feb. 2021), p. 1085 (cit. on pp. 54, 78, 89, 92).

[193] C. Song, D. Y. Park, P. L. Gehlbach, S. J. Park, and J. U. Kang. “Fiber-optic OCT sensor guided “SMART” micro-forceps for microsurgery”. In: *Biomedical Optics Express* 4.7 (July 2013), p. 1045 (cit. on p. 19).

[194] K. P. Soundararajan and T. Schultz. “Learning Probabilistic Transfer Functions: A Comparative Study of Classifiers”. In: *Computer Graphics Forum* 34.3 (2015), pp. 111–120 (cit. on p. 113).

[195] R. F. Spaide, J. G. Fujimoto, and N. K. Waheed. “Image Artifacts In Optical Coherence Tomography Angiograpy”. In: *Retina* 35.11 (Nov. 2015), pp. 2163–2180 (cit. on pp. 27, 28).

[196] J. T. Stout and P. J. Francis. “Surgical approaches to gene and stem cell therapy for retinal disease”. In: *Human Gene Therapy* 22.5 (2011), pp. 531–535 (cit. on pp. 6, 11, 14, 16, 17).

[197] P. Sullivan. *Vitreoretinal Surgery*. Eyelearning Ltd., 2014, p. 539 (cit. on pp. 11, 13, 16, 18).

[198] E. Sunden, A. Ynnerman, and T. Ropinski. “Image Plane Sweep Volume Illumination”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (Dec. 2011), pp. 2125–2134 (cit. on p. 32).

[199] E. Sunden and T. Ropinski. “Efficient volume illumination with multiple light sources through selective light updates”. In: *IEEE Pacific Visualization Symposium* 2015-July (2015), pp. 231–238 (cit. on pp. 32, 54, 95).

[200] E. A. Swanson, J. A. Izatt, M. R. Hee, et al. “In vivo retinal imaging by optical coherence tomography”. In: *Optics letters* 18.21 (1993), pp. 1864–1866 (cit. on p. 21).

[201] M. Szkulmowski and M. Wojtkowski. “Averaging techniques for OCT imaging”. In: *Optics Express* 21.8 (Apr. 2013), p. 9757 (cit. on p. 98).

[202] R. Sznitman, C. Becker, and P. Fua. “Fast Part-Based Classification for Instrument Detection in Minimally Invasive Surgery”. In: *Golland, P., Hata N., Barillot C., Hornegger J., Howe R. (eds) MICCAI 2014. LNCS, vol. 8673, pp. 692–699. Springer, Heidelberg (2014)* (cit. on p. 66).

[203] G. J. Tearney, B. E. Bouma, and J. G. Fujimoto. “High-speed phase- and group-delay scanning with a grating-based phase control delay line”. In: *Optics Letters* 22.23 (Dec. 1997), p. 1811 (cit. on p. 23).

[204] A. Tewari, O. Fried, J. Thies, et al. “State of the Art on Neural Rendering”. In: *Computer Graphics Forum* (2020) (cit. on p. 113).

[205] M. G. Tieger, K. Moussa, L. A. Kim, and D. Eliott. “The History of Visualization in Vitrectomy Surgery”. In: *International Ophthalmology Clinics* 60.1 (2020), pp. 1–15 (cit. on pp. 3–5).

[206] M. Tkalcic and J. Tasic. “Colour spaces: perceptual, historical and applicational background”. In: *The IEEE Region 8 EUROCON 2003. Computer as a Tool.* Vol. 1. IEEE, 2003, pp. 304–308 (cit. on p. 85).

[207] A. Tran, J. Weiss, S. Albarqouni, S. Faghi Roohi, and N. Navab. “Retinal Layer Segmentation Reformulated as OCT Language Processing”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 12265 LNCS. 2020, pp. 694–703 (cit. on pp. 54, 89).

[208] T. Ueta, T. Nakano, Y. Ida, N. Sugita, M. Mitsuishi, and Y. Tamaki. “Comparison of robot-assisted and manual retinal vessel microcannulation in an animal model”. In: *British Journal of Ophthalmology* 95.5 (May 2011), pp. 731–734 (cit. on p. 20).

[209] P. R. Van den Biesen, T. Berenschot, R. M. Verdaasdonk, H. Van Weelden, and D. Van Norren. “Endoillumination during vitrectomy and phototoxicity thresholds”. In: *British Journal of Ophthalmology* 84.12 (Dec. 2000), pp. 1372–1375 (cit. on p. 14).

[210] E. Vander Poorten, C. N. Riviere, J. J. Abbott, et al. *Robotic Retinal Surgery*. Elsevier Inc., 2020, pp. 627–672 (cit. on pp. 4, 15, 19, 20, 22, 35).

[211] C. Viehland, B. Keller, O. M. Carrasco-Zevallos, et al. “Enhanced volumetric visualization for real time 4D intraoperative ophthalmic swept-source OCT”. In: *Biomed Opt Express* 7.5 (2016), pp. 1815–1829 (cit. on pp. 81, 82, 86, 89, 90, 99).

[212] C. Viehland, A.-H. Z. Dhalla, J. D. Li, et al. “Real time volumetric intrasurgical optical coherence tomography with 4D visualization of surgical maneuvers (Conference Presentation)”. In: *Ophthalmic Technologies XXX*. Ed. by F. Manns, P. G. Söderberg, and A. Ho. SPIE, Mar. 2020, p. 19 (cit. on p. 26).

[213] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. “Image Quality Assessment: From Error Visibility to Structural Similarity”. In: *IEEE Transactions on Image Processing* 13.4 (Apr. 2004), pp. 600–612 (cit. on p. 122).

[214] W. Wei, R. Goldman, N. Simaan, H. Fine, and S. Chang. “Design and Theoretical Evaluation of Micro-Surgical Manipulators for Orbital Manipulation and Intraocular Dexterity”. In: *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE, Apr. 2007, pp. 3389–3395 (cit. on p. 20).

[215] J. Weiss, U. Eck, M. A. Nasseri, M. Maier, A. Eslami, and N. Navab. “Layer-Aware iOCT Volume Rendering for Retinal Surgery”. In: *Eurographics Workshop on Visual Computing for Biology and Medicine*. Ed. by B. Kozlíková, L. Linsen, P.-P. Vázquez, K. Lawonn, and R. G. Raidou. The Eurographics Association, 2019 (cit. on pp. 80, 97).

[216] J. Weiss, N. Rieke, M. Nasseri, M. Maier, A. Eslami, and N. Navab. “Fast 5DOF needle tracking in iOCT”. In: *International Journal of Computer Assisted Radiology and Surgery* 13.6 (2018) (cit. on p. 65).

[217] J. Weiss and N. Navab. “Deep Direct Volume Rendering: Learning Visual Feature Mappings From Exemplary Images”. In: *arXiv preprint* (2021). arXiv: 2106.05429 (cit. on p. 109).

[218] J. Weiss, N. Rieke, M. Maier, C. P. Lohmann, N. Navab, and A. Eslami. “Injection Assistance via Surgical Needle Guidance using Microscope-Integrated OCT (MI-OCT)”. In: *Investigative Ophthalmology & Visual Science July 2018* 59 (2018) (cit. on p. 65).

[219] J. Weiss, M. Sommersperger, A. Nasseri, A. Eslami, U. Eck, and N. Navab. *Processing-Aware Real-Time Rendering for Optimized Tissue Visualization in Intraoperative 4D OCT*. Vol. 12265 LNCS. Springer International Publishing, 2020, pp. 267–276 (cit. on p. 98).

[220] S. Weiss, M. Chu, N. Thuerey, and R. Westermann. “Volumetric Isosurface Rendering with Deep Learning-Based Super-Resolution”. In: *IEEE Transactions on Visualization and Computer Graphics* (2019), pp. 1–1. arXiv: 1906.06520 (cit. on p. 113).

[221] S. Weiss and R. Westermann. “Differentiable Direct Volume Rendering”. In: *arXiv preprint* (July 2021). arXiv: 2107.12672 (cit. on p. 113).

[222] M. Weisser. “RASIS Technologische Gestaltung eines Chirurgiesystems für die Augenheilkunde”. PhD thesis. Technical University of Munich, 2019 (cit. on p. 38).

[223] T. S. Wells, S. Yang, R. A. MacLachlan, J. T. Handa, P. Gehlbach, and C. Riviere. “Comparison of baseline tremor under various microsurgical conditions”. In: *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*. NIH Public Access, 2013, pp. 1482–1487 (cit. on pp. 14, 18).

[224] V. Westphal, A. Rollins, S. Radhakrishnan, and J. Izatt. “Correction of geometric and refractive image distortions in optical coherence tomography applying Fermat’s principle”. In: *Optics Express* 10.9 (May 2002), p. 397 (cit. on p. 29).

[225] J. R. Wilkins, C. A. Puliafito, M. R. Hee, et al. “Characterization of epiretinal membranes using optical coherence tomography”. In: *Ophthalmology* 103.12 (Dec. 1996), pp. 2142–2151 (cit. on p. 14).

[226] J. T. Wilson, M. J. Gerber, S. W. Prince, et al. “Intraocular robotic interventional surgical system (IRISS): Mechanical design, evaluation, and master-slave manipulation”. In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 14.1 (Feb. 2018), e1842 (cit. on p. 20).

[227] W. L. Wong, X. Su, X. Li, et al. “Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis”. In: *The Lancet Global Health* 2.2 (Feb. 2014), e106–e116 (cit. on p. 6).

[228] J. Xiao, Q. Wu, D. Sun, C. He, and Y. Chen. “Classifications and Functions of Vitreoretinal Surgery Assisted Robots-A Review of the State of the Art”. In: *Proceedings - 2019 International Conference on Intelligent Transportation, Big Data and Smart City, ICITBS 2019* (2019), pp. 474–484 (cit. on pp. 6, 18).

[229] S. Yang, R. A. MacLachlan, and C. N. Riviere. “Manipulator design and operation of a six-degree-of-freedom handheld tremor-canceling microsurgical instrument”. In: *IEEE/ASME Transactions on Mechatronics* 20.2 (2015), pp. 761–772 (cit. on p. 19).

[230] S. Yang, J. N. Martel, L. A. Lobes, and C. N. Riviere. “Techniques for robot-aided intraocular surgery using monocular vision”. In: *International Journal of Robotics Research* 37.8 (2018), pp. 931–952 (cit. on p. 35).

[231] M. Yip and N. Das. “Robot Autonomy for Surgery”. In: *The Encyclopedia of Medical Robotics*. World Scientific, Oct. 2018. Chap. 10, pp. 281–313. arXiv: 1707.03080 (cit. on p. 37).

[232] T. You, C. X. Huang, S. Chen, et al. “CASE REPORT: Extreme Positioning for Retinal Surgery in Advanced Kyphosis”. In: *Retinal Cases & Brief Reports* 9.3 (2015), pp. 218–219 (cit. on p. 11).

[233] D.-Y. Yu, S. Cringle, and I. Constable. “Robotic ocular ultramicrosurgery”. In: *Australian and New Zealand Journal of Ophthalmology* 26 (May 1998), S6–S8 (cit. on p. 20).

[234] A. Zhang, Q. Zhang, Y. Huang, Z. Zhong, and R. K. Wang. “Multifunctional 1050 nm spectral domain OCT system at 147 kHz for posterior eye imaging”. In: *Sovremennye Tehnologii v Medicine* 7.1 (2015), pp. 7–12 (cit. on p. 23).

[235] Q. Zhang, R. Eagleson, and T. M. Peters. “Volume Visualization: A Technical Overview with a Focus on Medical Applications”. In: *Journal of Digital Imaging* 24.4 (Aug. 2011), pp. 640–664 (cit. on pp. 26, 92).

[236] R. Zhang, J. Y. Zhu, P. Isola, et al. “Real-time user-guided image colorization with learned deep priors”. In: *ACM Transactions on Graphics*. Vol. 36. 4. 2017. arXiv: 1705.02999v1 (cit. on pp. 122, 133).

[237] D. Zhou, S. Kimura, H. Takeyama, et al. “Eye Explorer: A robotic endoscope holder for eye surgery”. In: *International Journal of Medical Robotics and Computer Assisted Surgery* March (2020), pp. 1–13 (cit. on p. 21).

[238] M. Zhou, H. Roodaki, A. Eslami, et al. “Needle Segmentation in Volumetric Optical Coherence Tomography Images for Ophthalmic Microsurgery”. In: *Applied Sciences* 7.8 (July 2017), p. 748 (cit. on pp. 35, 67).

[239] M. Zhou, X. Wang, J. Weiss, et al. “Needle localization for robot-assisted subretinal injection based on deep learning”. In: *Proceedings - IEEE International Conference on Robotics and Automation* 2019-May (2019), pp. 8727–8732 (cit. on pp. 35, 54, 58, 78).

[240] M. Zhou, Q. Yu, S. Mahov, et al. “Towards Robotic-assisted Subretinal Injection: A Hybrid Parallel-Serial Robot System Design and Preliminary Evaluation”. In: *IEEE Transactions on Industrial Electronics* 0046.c (2019), p. 1 (cit. on p. 38).

[241] M. Zhou, Q. Yu, S. Mahov, et al. “Towards Robotic-assisted Subretinal Injection: A Hybrid Parallel-Serial Robot System Design and Preliminary Evaluation”. In: *IEEE Transactions on Industrial Electronics* 0046.c (2019), pp. 1–1 (cit. on p. 98).

[242] C. Zuo, L. Mi, S. Yang, X. Guo, H. Xiao, and X. Liu. “The linear artifact in enhanced depth imaging spectral domain optical coherence tomography”. In: *Scientific Reports* 7.1 (Dec. 2017), p. 8464 (cit. on p. 27).

Online Resources

[243] American Academy of Ophthalmology. *Spark unveils $850,000 price tag for Luxturna*. 2018. URL: https://www.aao.org/headline/spark-unveils-850-000-price-tag-luxturna (visited on Oct. 2, 2021) (cit. on p. 6).

[244] P. Bracha and T. A. Ciulla. *Stem Cell Therapy in Retinal Disease*. 2018. URL: https://www.reviewofophthalmology.com/article/stem-cell-therapy-in-retinal-disease (visited on Nov. 6, 2020) (cit. on p. 16).

[245] B. Ley. *Diameter of a Human Hair*. 1999. URL: https://hypertextbook.com/facts/1999/BrianLey.shtml (visited on Nov. 4, 2020) (cit. on p. 14).

[246] NIH | National Eye Institute. *Macular Pucker*. URL: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/macular-pucker (visited on Nov. 5, 2020) (cit. on p. 15).

[247] Preceyes B. V. *PRECEYES Surgical System – Preceyes BV*. URL: http://www.preceyes.nl/preceyes-surgical-system/ (visited on Nov. 11, 2020) (cit. on pp. 4, 5, 18, 20).

[248] Zeiss AG. *OPMI LUMERA 700 from ZEISS AG*. URL: https://www.zeiss.com/meditec/int/product-portfolio/surgical-microscopes/ophthalmic-microscopes/opmi-lumera-700.html (visited on June 6, 2019) (cit. on p. 89).

List of Figures

List of Tables