Martin Huber

Data-driven Robotic Endoscope Automation

School of Biomedical Engineering & Imaging Sciences
King's College London

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

30th March 2024

Technical Supervisor
Prof. Christos Bergeles

Technical Co-Supervisor
Prof. Tom Vercauteren

Für meine Eltern, Margarete und Walter.

Nur durch eure unerschütterliche Hilfe, euer Vertrauen, euer Wissen, euer Vorbild, euren Zumut, eure Zeit, Liebe und eure Stärke, war diese Doktorarbeit möglich. Diese Arbeit widme ich euch. Danke dafür, mir diesen Weg ermöglicht zu haben.
Martin.

The best way to predict the future is to invent it.

— Steve Jobs [Isaacson 2011]

I, Martin Huber, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Automation is increasingly integrating into our daily routines, with autonomous vehicles, vacuum cleaners, drones, and industrial robots exhibiting notable advancements in autonomy. These developments owe much to the significant boost in parallel computing capabilities and the widespread use of neural networks. However, in fields such as surgery and robotic surgery, automation progress lags behind other sectors. Apart from challenges posed by irregular anatomy, regulatory constraints, and ethical considerations, this thesis contends that technical barriers also impede further advancements. As indicated by the thesis title, *Data-driven Robotic Endoscope Automation*, robotic endoscope automation holds immediate relevance in this context due to its relative simplicity. Traditional rule-based methods, however, oversimplify camera motion and assume a tool-following pattern. The adoption of modern data-driven approaches could facilitate the learning of more intricate control policies. Two primary technical obstacles hinder progress in this area: the absence of state-action pairs in public datasets, rendering the application of modern imitation learning (IL) techniques to actual clinical data impossible, and the lack of clear pathways toward automation in currently non-robotic surgeries.

In this thesis, we address the hurdles through developing methods that extract camera motion control policies from surgeon-held endoscopes, as well as methods for executing the policies on serial manipulators. First, we introduce a novel marker-free unified hand-eye calibration procedure that allows for precise robot localization from RGB-D images in a clinical setting whilst keeping the clinical workflow unaltered. This is a prerequisite to spatial awareness and to control robots autonomously in a surgical theater. We verify the proposed method during an in-vivo experiment and demonstrate successful application. Second, we derive a visual servo for image-based control that does not require any explicit tool and camera position nor any explicit depth information. The proposed method works regardless of the relative patient coordinate frame and is safe for clinical use by design. We deploy the approach in a semi-autonomous scenario. Third, we find efficient means for extracting state-action pairs from retrospective videos of laparoscopic surgery videos (surgeon-held or robotic) through a novel data augmentation method in a supervised learning

procedure. The proposed method is robust enough that it transfers from videos of robotic laparoscopic surgery to classical laparoscopic interventions. Fourth, we introduce a fully self-supervised pipeline to learn to predict actions from the observed states through generating pseudo-ground truth via the action extraction. Finally, we contribute significantly to robot driver infrastructure and ecosystem embedding in an attempt to make advancements accessible to a greater community and in return benefit from community contributed developments for accelerated research and deployment.

Impact Statement

The research outlined in this PhD thesis holds the promise of significant global impact across societal, clinical, and academic realms. As we cautiously assess its early impact, we anticipate its future contributions to be widely uptaken.

Societal Impact Given its primary focus on automation, this thesis holds potential for societal impact, encompassing both advantageous and detrimental implications. On the positive spectrum, automation has the capacity to sustain healthcare accessibility within a society confronting labor shortages, potentially reducing associated costs. Moreover, it could facilitate the redirection of human efforts towards more enriching endeavors. Conversely, the advent of automation may precipitate job displacement or compel individuals into roles perceived as less fulfilling.

Clinical Impact This research has the potential to expedite clinical delivery by mitigating reliance on the presence of clinical personnel. As this thesis delves into the preliminary stages of camera motion automation, it offers a foundation for extrapolating insights into automating a broader spectrum of tasks in subsequent endeavors. Considering potential economic motivations, it is imperative to ensure that newly devised systems remain within current cost ranges. It is essential to acknowledge that, akin to surgeons, automation carries the risk of patient harm in the event of errors. However, the methodologies developed within this thesis shed light on potential strategies to mitigate such risks even in automated settings.

Academic Impact The academic significance of this research primarily lies in the proposed self-supervised learning methodologies for camera motion automation, ensuring safe implementation in clinical settings. These innovations have the potential to catalyze the emergence of new research domains while steering existing fields towards unexplored avenues. Notably, for the first time, this study investigates automation prospects on a comprehensive scale using data from large-scale videos of laparoscopic surgeries. Although it is premature to forecast additional advancements in automation stemming from this thesis, we are optimistic. Future research could for example build on the proposed self-supervised training scheme

with added and improved building blocks for higher levels of autonomy. The idea of bridging the human-robot embodiment-gap through optimal control could further be extended through more sophisticated approaches. Finally, research could pick up the in-vivo surgery data with eye-to-hand registrations to solve marker-free registration despite draping and occlusions.

Acknowledgments

First and foremost, I would like to express my dearest gratitude to both my supervisors Christos and Tom. Thank you for giving me the chance to come work with you at King's College London (KCL), in this vibrant environment. It has been a once in a lifetime opportunity to come here. It has not always been the straightest route. We worked through COVID together, you kept the spirits up, despite your hardship of working extra hours and schooling your children. You were forgiving when things went the wrong way, and encouraged me to keep pushing. Thank you for being available day and night, during holidays and weekends, for proofreading every paper, for taking the time to meet regularly, for hinting me in the right direction and listening to stupid ideas, thank you for creating opportunities, letting me go to places, and exposing me to all your network. I cannot thank you enough. It is only for your brilliance, dedication, and support that this thesis exists. I hope this is just the start and that we can continue on this journey together.

Next, I would like to thank the thesis progression committee (TPC): Adelaide, Kawal, Hongbin, Prashant. Your feedback helped shape the research targets immensely. I would like to express a special thanks to Adelaide, the TPC chair, for always being available and responding to requests promptly.

Thank you Sébastien, Mark, and Jorge for creating these beautiful groups that contributed significantly to the creation of this laboratory, office, and the amazing integration into the hospital.

Furthermore, I would like to thank the Robotics and Vision in Medicine (RViM) team for insightful discussions, positive feedback, help with the thesis, as well as experiments, and many good memories: John, Harris, Carlo, Anestis, Hans, Brandon, Zicong, Jeremy, Mirroyal, Jeref, Maleeha, Claudio, Gongyu, Soutiris, Zisos, Theo, and Hadi. A special thanks to Lyndon, for inviting me to many eye surgeries, for sharing your knowledge, for your growth mindset, and your pragmatism, many of which have been incredible learnings.

Thank you Chayanin, Ross. Thank you for being great friends and making me feel at home in London. Thank you for all your technical recommendations, knowledgeable

feedback, for thinking through ideas, setting things up, for maintaining a healthy working environment and for not taking things too seriously despite the exhausting working hours of a researcher that is. Only for your help, this submission turned into what it is today. I hope we can continue being friends and travel together in the future.

I would especially like to thank Christopher, Huanyu, Charlie, Dan, and Luis for extremely fruitful collaborations and plenty of technical feedback, much of which got incorporated into this work. It has been an absolute gift to get to work with you and I hope we get to work together again on another occasion in the future. I would love to learn a lot more from you and hear about your thoughts.

I would further like to thank Alejandro, Julien, Nicholas, and Yang for inviting me to a robotic surgery training, which provided some final insights to this thesis. I would especially like to thank Alejandro for being a great friend. I hope we get to go on many more runs together, maybe have Lucas help keep us accountable, whom I'd also like to thank.

I would like to thank the team in the optics lab and their help around the equipment, and just great input, and insights overall: Michael, Philip, Mirek, Theo, and Yijing. It has been incredible to get to bump into you on a daily basis, and I hope we will keep in touch. Thank you also, Yue, for proofreading.

Thank you also, Samuel, Lilly, Mikel, and Viktor, my roommates throughout this journey. Thank you for being good roommates, for valuable discussions that helped inspire this research during many occasions.

Thank you to Theo, Maxence, Tiarna, Virginia, Pedro, Mark, Aaron, Remi, Tosin, Stephen, and Reuben. Thank you for being available when no one was around. Thank you for teaching me how to submit jobs to the DGX, for your great inspiration, and for organizing Tommies' Social.

Thank you to the KCL staff, for keeping things running. I would especially like to thank the IT team for supporting the graphics processing unit (GPU) cluster and providing the bouncer server: Andrew and Davide. Thank you also to Laurence for always helping with purchase orders swiftly. Thank you, Irina for being a good friend and for keeping us in the pipe. I would further like to thank David for getting the laboratories up and running promptly during COVID, and for helpful discussions. My gratitude further goes to the SIE managers: Valentina, Marty, Duane, and Gayathri. Thank you for always helping with printing, moving, custom building screws and metal blocks, and for finding equipment. Without you, none of

the experiments would have been possible. I would further like to thank the security personnel and the cleaning lady. Thank you for letting me into Becket House and the SIE laboratories day and night. Thank you further for keeping my desk in an orderly state.

I am deeply thankful for meeting Konrad, for him giving me the opportunity to work in the US during an internship. Thank you for all your hard work, your mentorship, and for making me feel at home. Some of the learnings are now ingrained in this work, too. I would further like to thank his team for their continued help on-site, for playing volleyball together, going to hikes, and many more occasions: Eva (Yuqing), Olivier, Bryan, Ahmed, Alisha, Alan. Thank you also, Kathleen, for all your help.

I am very thankful for the collaborations with the entire functionally accurate robotic surgery (FAROS) team. Pulling a project of this scale off has not always been easy and required heavy lifting work from many participants. Thank you to the team from Leuven: Manu, Ayoob, Maikel, Ruixuan, and Kaat, for their input and support, especially for letting us use their robot. Thank you to the team from Zurich: Philipp, Fabio, Frederic, Nicola, and Aidana. A special thanks to Fabio, for having me stay on his couch during one of the integration weeks and for being a good friend. Thank you also to Frederic, for the joint work on robot drivers. Thank you to the team from Paris: Guillaume, Antoine, Jimmy, Francois, Saman, Lilian, Thibault, and Ellie. Many of the experiments that contributed to this thesis would not have been possible without your work.

Nun auf meiner Muttersprache. Danke Thomas, Christian und Matthias dafür gute Brüder zu sein. Euer gutes Vorbild waren mir stets eine Hilfe dabei den rechten Weg zu finden. Thomas durch seine Entschlossenheit und Umsetzungsfähigkeit, sowie der Liebe zu Büchern. Christian und Matthias durch ihren Kampfgeist im Sport und der gesunden Ernährung. Ohne euer Zusprechen, eure Hilfe von klein auf bei Matheaufgaben, eure Inspiration Ziele zu verfolgen, wäre diese Arbeit wohl nicht möglich gewesen. Was in Zukunft auf uns zukommen wird, wird nicht leicht, aber auch hierfür werden wir einen Weg finden und ich hoffe, dass wir dabei die gemeinsame Freude erhalten können. Auf das wir noch viele schöne Sommerabende beim Grillen verbringen können. Außerdem möchte ich meinem Paten Eberhard für seine langjährige Unterstützung danken, sowie seiner Ermutigung zum akademischen Weg.

Danke auch an Guan Teck und Philipp, dafür gute Freunde zu sein. Wär hätte gedacht, dass diese Arbeit mit einem Jobs Zitat startet, aber nun gibt es wohl kein Zurück mehr. Ein besonderer Dank an Guan Teck, lange ist es her, dass du mir

dabei geholfen hast computed tomography (CT) Binaries in C++ einzulesen, aber durch deine Hilfe bei diesem ersten Schritt wurde der Rest ermöglicht.

Weitherhin möchte ich meinen ehemaligen Mitstudenten und Freunden aus Heidelberg danken. Danke Lucas und Lucas, unser ständiger Ideenaustaucht, insbesondere durch unseren Journal Club während COVID, aber auch im Allgemeinen, hat einige Einsichten dieser Arbeit geprägt. Danke auch dafür, gute Freunde zu sein. Dein Besuch in London zu Beginn der Arbeit, Lucas, war ein wichtiges Erlebnis und er hat mir sehr dabei geholfen meinen Weg zu finden. Überhaupt wäre das Physik Studium ohne dich vermutlich nicht so glatt gelaufen, wie es das tat. Danke auch an Markus und dafür, dass ich bei dir zu Ende der Masterarbeit wohnen durfte, was das alles irgendwie in Gang gebracht hat. Danke auch an deine Oma.

Ich möchte außerdem einigen Lehrern danken, die mich besonders geprägt haben. Danke Herr Blomeier, Herr Steinhäuser und Frau Schmickler für Ihren exzellenten Unterricht. Danke auch an meine ehemaligen Fußballtrainer Stefan, Jens und Kai. Danke weiterhin an meine Bachelor und Master Betreuer aus Heidelberg Marc Kachelrieß und Katja Mombauer, danke für Alles was ich in euren Gruppen lernen durfte und die Empfehlung, die es ermöglicht hat nach London zu kommen.

Zu guter Letzt möchte ich dem Team der International Feedaz für ihren anhaltenden Hype danken, der die ein oder andere dunkle Nacht erleuchtet hat. Von Top bis Bot: Lukas, Hendrik, Viktor und Jungle Diff Dennis.

List of Publications

First Author

International Peer-reviewed Journals

- Martin Huber, Sébastien Ourselin, Christos Bergeles, and Tom Vercauteren: *Deep Homography Estimation in Dynamic Surgical Scenes for Laparoscopic Camera Motion Extraction*, [Huber 2022].

Under Review

- Martin Huber, Christopher E. Mower, Sébastien Ourselin, Tom Vercauteren, and Christos Bergeles: *LBR-Stack: ROS 2 and Python Integration of KUKA FRI for Med and IIWA Robots*, [Huber 2023a].

- Martin Huber, Huanyu Tian, Christopher E. Mower, Charlie Budd, Samuel Joutard, Ayoob Davoodi, Saman Vafadar, Antoine Harle, Emmanuel Vander Poorten, Guillaume Morel, Christos Bergeles, and Tom Vercauteren: *Hydra: Stereo Imaging Approach to Unified Vision-based Robot Calibration*, not public yet.

International Peer-reviewed Conferences with published Proceedings

- Martin Huber, John Bason Mitchell, Ross Henry, Sébastien Ourselin, Tom Vercauteren, and Christos Bergeles: *Homography-based Visual Servoing with Remote Center of Motion for Semi-autonomous Robotic Endoscope Manipulation*, [Huber 2021].

- Martin Huber, Sébastien Ourselin, Christos Bergeles, and Tom Vercauteren: *Deep Homography Prediction for Endoscopic Camera Motion Imitation Learning*, [Huber 2023b].

Co-author

International Peer-reviewed Journals

- Charlie Budd, Luis C. Garcia-Peraza Herrera, Martin Huber, Sébastien Ourselin, and Tom Vercauteren: *Rapid and robust endoscopic content area estimation: a lean GPU-based pipeline and curated benchmark dataset*, [Budd 2023a].

International Peer-reviewed Conferences with published Proceedings

- Charlie Budd, Jianrong Qiu, Oscar MacCormac, Martin Huber, Christopher E. Mower, Mirek Janatka, Théo Trotouin, Jonathan Shapey, Mads S. Bergholt, and Tom Vercauteren: *Deep Reinforcement Learning Based System for Intraoperative Hyperspectral Video Autofocusing*, [Budd 2023b].
- Huanyu Tian, Martin Huber, Christopher E. Mower, Zhe Han, Changsheng Li, Xingguang Duan, and Christos Bergeles: *Excitation Trajectory Optimization for Dynamic Parameter Identification Using Virtual Constraints in Hands-on Robotic System*, [Tian 2024].
- Christopher E. Mower, Martin Huber, Huanyu Tian, Ayoob Davoodi, Emmanuel Vander Poorten, Tom Vercauteren, and Christos Bergeles: *Vision and Contact based Optimal Control for Autonomous Trocar Docking*, [Mower 2023a].

Pre-print Articles

- Imanol Luengo, Maria Grammatikopoulou, Rahim Mohammadi, Chris Walsh, Chinedu Innocent Nwoye, Deepak Alapatt, Nicolas Padoy, Zhen-Liang Ni, Chen-Chen Fan, Gui-Bin Bian, Zeng-Guang Hou, Heonjin Ha, Jiacheng Wang, Haojie Wang, Dong Guo, Lu Wang, Guotai Wang, Mobarakol Islam, Bharat Giddwani, Ren Hongliang, Theodoros Pissas, Claudio Ravasio, Martin Huber, Jeremy Birch, Joan M.Nunez Do Rio, Lyndon da Cruz, Christos Bergeles, Hongyu Chen, Fucang Jia, Nikhil KumarTomar, Debesh Jha, Michael A. Riegler, Pal Halvorsen, Sophia Bano, Uddhav Vaghela, Jianyuan Hong, Haili Ye, Feihong Huang, Da-Han Wang, and Danail Stoyanov: *2020 cataracts semantic segmentation challenge*, [Luengo 2022]. First in two out of three challenges.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Table of Contents

1.1 Foreword

This thesis investigates data-driven endoscopic camera motion automation by means of a robot and therefore falls into the realm of robot assisted surgery (RAS). Particular emphasis is hereby put on automating currently non-robotic surgeries. Through research on robotics and computer vision, this work thus aims to find novel ways to support clinical staff in their surgical work by alleviating the burden of this unfulfilling and repeatable task.

In RAS, it is commonly acknowledged that automation will proceed stepwise. From no autonomy to full automation, these steps are often categorized into six stages [Yang 2017; Fosch-Villaronga 2021]:

1. No autonomy: Surgeon is in full charge of the robot.
2. Robot assistance: Robot constrains motion and corrects surgeon.
3. Task autonomy: Robot executes specific tasks autonomously under human supervision.
4. Conditional autonomy: Robot plans and executes tasks under human approval.
5. High autonomy: Conditional autonomy without approval but with human intervention.
6. Full automation: Robot performs an entire surgery autonomously.

Aiming at fully imitating assistant surgeon's actions, this thesis therefore explores level five, high autonomy, which would require approval-free endoscopic camera motion execution with the possibility of human intervention. Although some works [Battaglia 2021] argue that the focus should not lie on automation, but enhancement, i.e. guiding a surgeon's intent, we adopt a futuristic sentiment and are in alignment with [Kitaguchi 2022], where the authors explain why camera motion automation will likely be achieved first. This futuristic notion follows previous surgical revolutions, from the advancement of open surgery to MIS, and the success of the da Vinci® surgical robot in RMIS, see also Fig. 1.1 and Fig. 1.2.

The ever evolving field of surgery has repeatedly demonstrated acceptance of new technology [Attanasio 2021], and we argue that automation is no exception. The growth of RMIS can clearly be considered an enabling factor for full automation, see also Fig. 1.1.

When compared to other domains, such as autonomous driving, service and house-

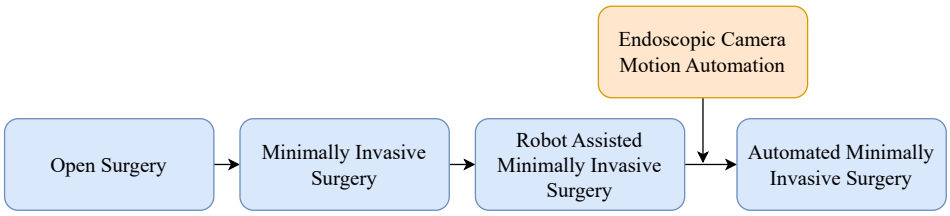


Figure 1.1: The gradual introduction of novel technology into the surgical field, one enabling the next. Endoscopic camera motion automation is likely to appear first towards full automation. Refers to Section 1.1.

hold robots, where robots already exert a significant amount autonomy that far exceeds RMIS, it becomes evident that advances in autonomy will ultimately make their way into the operating theater as well. Regulatory and economic hurdles cause delays in the adoption of novel technology by healthcare systems and hospital units. However, there might even be a future where RMIS will not only adopt automation techniques from related fields but make some contributions to general machine intelligence through its unique human-robot collaborative nature.

LeCun [2022] e.g. argued that machines will need to interact with the real world in order to become truly intelligent and, arguably, RMIS has grown to be the most advanced physical human-robot interaction (pHRI) domain with thousands of deployed robots.

As we dive deeper into level five autonomy endoscopic camera motion automation, this introductory chapter will provide a necessary clinical background for laparoscopy in Section 1.2, a sub-domain of endoscopy, including an in-depth explanation of laparoscopic cholecystectomy (minimally invasive gallbladder removal). Laparoscopic cholecystectomy is the most commonly performed laparoscopic procedure [Sheetz 2020], and can be considered relatively simple. Therefore, due to vast amounts of readily available data, it serves as the perfect test-bed for the methods that are derived as part of this thesis.

Next, we explain the rise of robot assisted laparoscopy in Section 1.3 and the potentials of automation that evolve with it. We highlight different commercial systems, their limitations, and propose innovative solutions. Identified limitations will include spatial awareness, refer Section 1.3.3.1, as most RMIS systems lack a world model, and, crucially, camera motion automation, for which a thorough review of related methods is provided in Section 1.5. Following the camera motion automation literature review, we will hypothesize an approach to near-term full automation of robotic laparoscopes. The proposed approach will revolve around a mixture of IL an classical control. It is grounded in successful automation in related domains

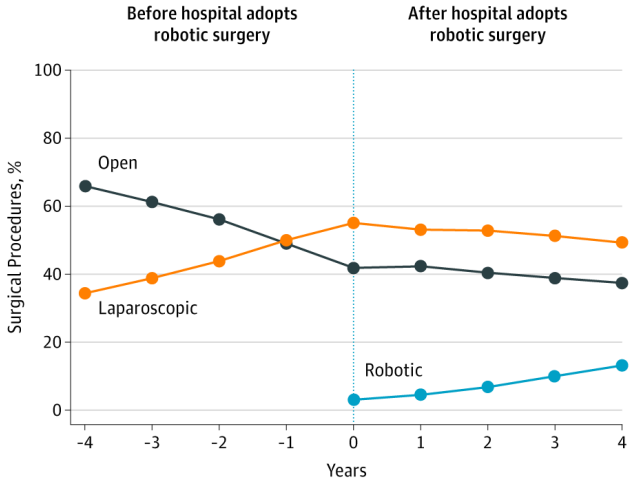


Figure 1.2: Increase of laparoscopy over open abdominal surgeries and robotic laparoscopy versus other methods, respectively. Data normalized to year zero, the introduction of robotic laparoscopy. The average use of robotic laparoscopy increased from 1.8% to 15.1%. For inguinal hernia repair, an increase in robotic surgery from 0.7% to 28.8% was found. A total of 169.404 cases over 73 hospitals in Michigan, United States, were investigated. Figure and data provided with courtesy of [Sheetz 2020]. Refers to Section 1.1.

and will be discussed in Section 1.6. Finally, Section 1.7 will provide an overview of the remaining thesis structure.

1.2 Laparoscopy

Endoscopy refers to the observation of internal parts by means of an endoscope [OED Online 2023]. The word endoscopy derives from the Greek by combining the prefix "endo" meaning "within" and the verb "skopein", "to view or observe" [Majumdar 1993]. In the surgical context, endoscopy refers to a MIS procedure with the goal of observing within the body using a rigid endoscope. Surgical endoscopy inside the abdominal (belly) or pelvic (hip) cavity is called laparoscopy. During laparoscopy, a rigid endoscope, called laparoscope in this clinical context, is inserted through small incisions for diagnostic or interventional purposes, see Fig. 1.3.

Compared to open surgery, laparoscopy leads to faster interventions (57.19 ± 10.13 min vs. 85.10 ± 15.18 min), less bleeding complications (2% vs 7% of interventions), shorter hospital stays (2.1 ± 1.1 days vs 4.4 ± 2.1 days), less infections, and less postoperative pain [Shi 2024]. These eminent advantages of laparoscopic over open techniques have led to their adoption since their introduction in the early 1980s, see Fig. 1.2. This trend is foreseen to continue and even accelerate as the major-

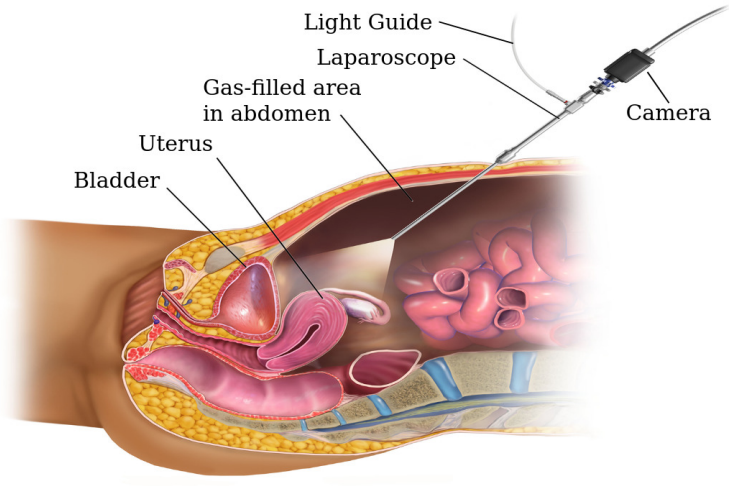


Figure 1.3: Illustration of a laparoscopic procedure. The laparoscope is inserted through a small incision in the patient's abdominal wall and provides a view of the surgical scene. To provide space, carbon dioxide (CO_2) is injected through a needle into the abdomen. Image provided with courtesy of [Blausencom 2014]. Refers to Section 1.2.

.

ity of surgical residents is nowadays trained on the minimally invasive procedure variants [John 2020].

Some types of laparoscopic procedures are cholecystectomy (gallbladder removal), appendectomy (appendix removal), inguinal hernia repair (i.e. leaking of intestines through the abdominal wall), colectomy (partial colon removal), nephrectomy (partial or complete kidney removal), prostatectomy (partial or complete prostate removal), adrenalectomy (removal of the adrenal gland), and gastrectomy (partial or full stomach removal) among others. Since laparoscopic cholecystectomy is carried out in high volumes and is additionally a relatively simple procedure, it is of special relevance to this work. It will thus be considered an example for explaining a laparoscopic surgery setup and procedure.

Cholecystectomy is the surgical removal of the gallbladder. The gallbladder serves as a reservoir for the liver-produced bile, a fat digesting fluid. Bile may harden and form gallstones, which can lead to inflammation and severe pain. Since the liver also releases bile directly into the digestive tract, shown in Fig. 1.5, the gallbladder may be removed when necessary. A cholecystectomy setup is depicted in Fig. 1.4.

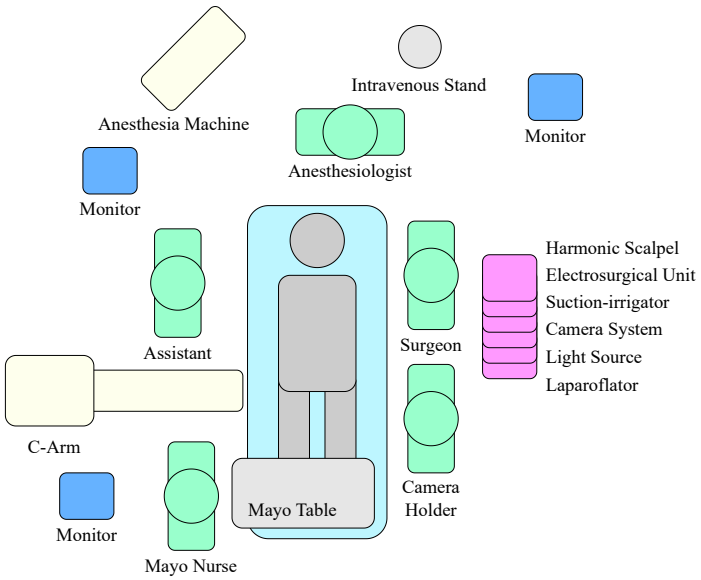


Figure 1.4: Typical setup for a cholecystectomy. Image adapted from [SAGES 2010]. Refers to Section 1.2.1.

1.2.1 Laparoscopic Cholecystectomy Setup

Even this relatively simple procedure requires multiple skilled practitioners. The surgeon is supported by an endoscope holder, a second assistant, an anesthesiologist, and a mayo nurse (for providing tools). The procedure commonly requires a harmonic scalpel, which cauterizes tissue and coagulates blood (i.e. causes blood clotting) through ultra-sonic waves. Due to lower cost for cauterization and coagulation, it may be preferable to use an electrosurgical unit (ESU), which generates different electrical waveforms and can be connected to most laparoscopic tools [Archana 2018]. Further devices include a suction-irrigator for removing body fluids from the surgical scene, as well as an endoscopic camera system, a light source, and multiple monitors. A laparoflator is used to insulflate the abdomen with CO_2 and to maintain a fixed pressure. An anesthesia machine and an intravenous stand are utilized to regulate the patient's conditions. Finally, a C-arm may visualise the bile duct to probe for possible injuries and leakage caused by the intervention [Cuschieri 1994]. The procedure is then referred to as cholangiography (x-ray visualization of the bile duct with contrast agent).

1.2.2 Laparoscopic Cholecystectomy Procedure

This paragraph explains laparoscopic cholecystectomy in simplified terms and vastly refers to [Majumder 2020]. To begin with, CO_2 is injected into the abdomen via a needle. The pressure is controlled through the laparoflator, see Fig. 1.4. Next,

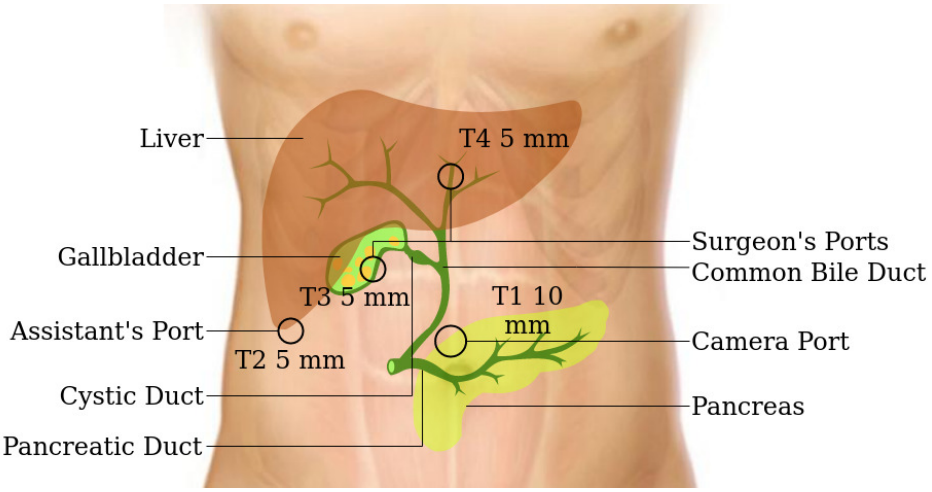


Figure 1.5: Common cholecystectomy incisions. Several incisions are made for the trocars T1-T4. Image with courtesy of [Majumder 2020] and modified to include port descriptions and organs. Refers to Section 1.2.2

four small incisions are made for trocars T1-T4, which serve as entry point through the abdominal wall, shown in Fig. 1.5. A laparoscope is inserted through T1 to grant an adequate view of the surgical scene. A grasper is inserted through T2 and used by the assistant to elevate the gallbladder. Trocars T3 and T4 are used by the surgeon to perform the cholecystectomy. It is important to identify anatomical landmarks before other steps are attempted. A good exposure of the surgical area is achieved through adequate patient and port positioning [Gupta 2023]. Following the preparatory steps, the surgery can broadly be partitioned into six steps. These steps will be explained in the following.

1.2.2.1 Step 1: Dissection of the Hepatocystic Triangle

The goal is to expose the hepatocystic triangle [Mischinger 2020], see Fig. 1.6. The gallbladder is gently pulled upward over the liver and its neck is pulled downward to expose its different parts, as can be seen in Fig. 1.5. Swollen gallbladders are precautiously decompressed with a needle to prevent perforation with leakage of bile and gallstones. Potential adhesions (scarred connections to other organs) are carefully separated using cautery and regular tools, avoiding the area near the duodenum (beginning of the small intestine). Next, the hepatocystic triangle is dissected by carefully removing fibrous and fatty tissue. It is of utmost importance that no tubular structure may be damaged until cystic duct and cystic artery are identified [Mischinger 2020].

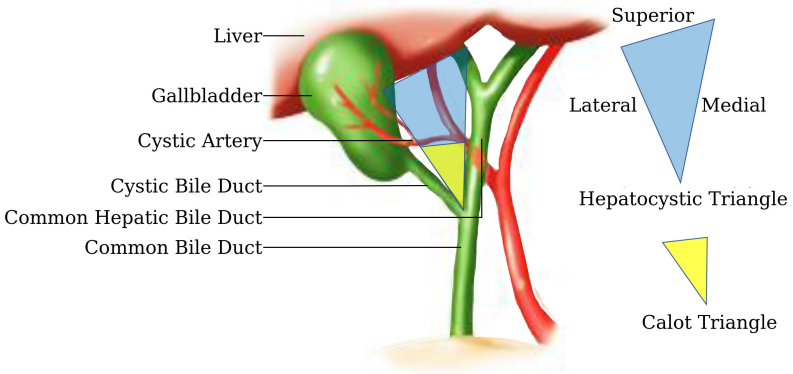


Figure 1.6: The bile duct (green), together with liver and gallbladder form the hepatocystic triangle, which is often covered in fat tissue. Image with courtesy of [Mischinger 2020] and updated font. Refers to Section 1.2.2.

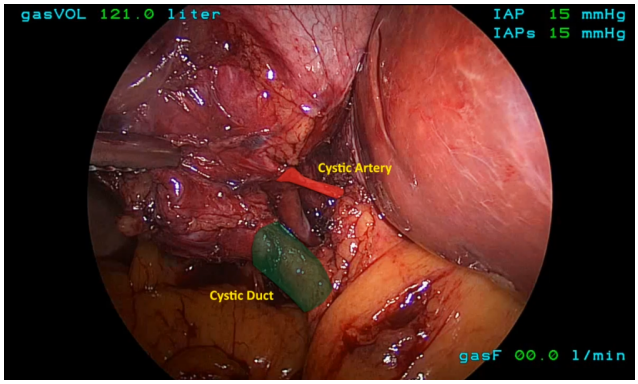


Figure 1.7: The critical view of safety. The cystic artery is indicated in red, the cystic duct in green. Image provided with courtesy of [Majumder 2020]. Refers to Section 1.2.2.

1.2.2.2 Step 2: Establishing the Critical View of Safety

The critical view of safety defines a state that is achieved through the hepatocystic triangle dissection, it is shown in Fig. 1.7. No tubular structure may be damaged prior to achieving the critical view of safety. The critical view of safety is defined by three conditions. First, the hepatocystic triangle is cleared of all fat and fibrous tissue. The common bile duct and the common hepatic bile duct are identified but not exposed. Second, the lower third of the gallbladder is separated from the liver. Third, only two structures are identified entering the gallbladder, the cystic duct and the cystic artery. The surgery is halted in this stage and reflected upon. Potential anatomical variations are discussed.

1.2.2.3 Step 3: Clipping and Cutting the Cystic Artery

After establishing the critical view of safety, the next step is to separate the gallbladder from the cystic artery. Therefore, the csytic artery is clipped twice proximally (i.e. the side of the artery that stays inside the body) and once distally (i.e. on the side of the to be removed gallbladder). The distal clip should be attached close to the neck of the gallbladder to leave sufficient space for cutting. Next, the artery is cut with hook scissors, leaving some space at the proximal end to prevent the clip from detaching.

1.2.2.4 Step 4: Operative Cholangiography and Cutting the Cystic Duct

Post cutting the cystic artery, it is sometimes preferred to perform cholangiography, using the C-arm X-ray, which is shown in Fig. 1.4. This is to verify a functioning bile tree. First, the cystic duct is clipped distally, and then incised partially with hook scissors. Next, the common bile duct is flushed with saline and a contrast agent is injected. Fluoroscopy should reveal free flow of the contrast agent into the common hepatic bile duct, its left and right branches into the liver, as well as free flow into the duodenum via the common bile duct. Once satisfactory flow is observed, the cystic duct is clipped proximally and cut as the cystic artery in step three.

1.2.2.5 Step 5: Separating Gallbladder from Liver

After cutting the cystic duct, the gallbladder is carefully dissected from the liver bed, commonly using an L-hook with monopolar energy. Care must be taken to prevent liver bed injury as this may cause bleeding and or bile leakage from superficial ducts. Entry into the gallbladder should be avoided as this may cause bile and bile stone leakage, which, however, can be removed with a suction irrigator, as shown in Fig. 1.4. In complex conditions with gallbladder inflammation, an ultrasound-driven harmonic scalpel may be used to perform ultrasonic coagulation to maintain hemostasis (blood thickening). Finally, the liver bed is once more inspected before separating the gallbladder fully. The surgical side is once more irrigated and cleaned of any blood and bile. The gallbladder is then put into a specimen bag.

1.2.2.6 Step 6: Removing Specimen and Closing

Once in a bag, the gallbladder is removed through the T1 port, refer to Fig. 1.5. This may require widening of the opening in the presence of larger stones. The

ports T1-T4 are being vented to remove residual CO_2. Following that, the skin and fascia at the extraction site are closed with sutures.

1.3 Robot Assisted Laparoscopy

Having a good understanding of laparoscopic surgery from the previous Section 1.2, and more specifically, laparoscopic cholecystectomy, we will next discuss robot assisted laparoscopic surgery. We will initially explain the, at first glance, paradoxical rise of robot assisted laparoscopy in Section 1.3.1. Current commercial systems will be analyzed in Section 1.3.2, to better understand where the field might be headed. Finally, the enhancement of current systems will be discussed in Section 1.3.3.

1.3.1 The Rise of Robot Assisted Laparoscopy

According to many, the future of surgery is robotic [Bakalar 2021]. This trend was already exposed in Section 1.1, Fig. 1.2, where Sheetz et al. [2020] found an average increase in robotic laparoscopy from 1.8% to 15.1% post robotic laparoscopy introduction. For cholecystectomy, the most common procedure, an increase from 2.5% to 7.5% was found. Somewhat paradoxically, (non-robotic) laparoscopic cholecystectomy is already a routine procedure with very low complication rates, Thapar et al. [2023] e.g. report a 0.2% 30 days mortality rate in India. Hence, questions about benefits of robotic laparoscopy may rightfully be raised.

Patient Side Aspects More than twenty years have passed since the introduction of the da Vinci® surgical robot by Intuitive (Fig. 1.9a), which was launched in 1999, and many review studies compared classical vs. robotic laparoscopies ever since to provide evidence-based care. The majority of studies suggests that no significant advantages exist for patients. Despite increased cost and longer intervention times, these studies report equal complication, mortality, and conversion rates (i.e. conversion to open surgery), as well as equal post operative stay duration [Kawka 2023; Csirzó 2023]. Some studies, however, also highlight underrepresented patient-reported outcome measures [Kawka 2023] (e.g. post-operative pain, return to function).

Surgeon Side Aspects Undoubtedly, the market for RAS, which is currently dominated by Intuitive Surgical, Inc., is sustainably growing.

The reasons for this steady growth, as per above and to the best current knowledge, is not driven by patient benefits. Instead, the growth can be linked to surgeon

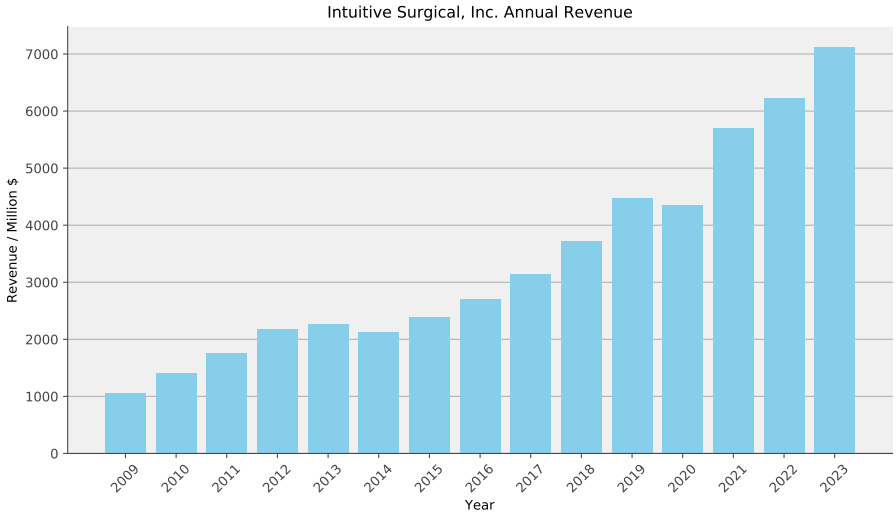


Figure 1.8: Annual revenue of Intuitive Surgical, Inc. from 2009 to 2023. The company has achieved sustained growth throughout the years. Data obtained from [Macrotrends 2024].

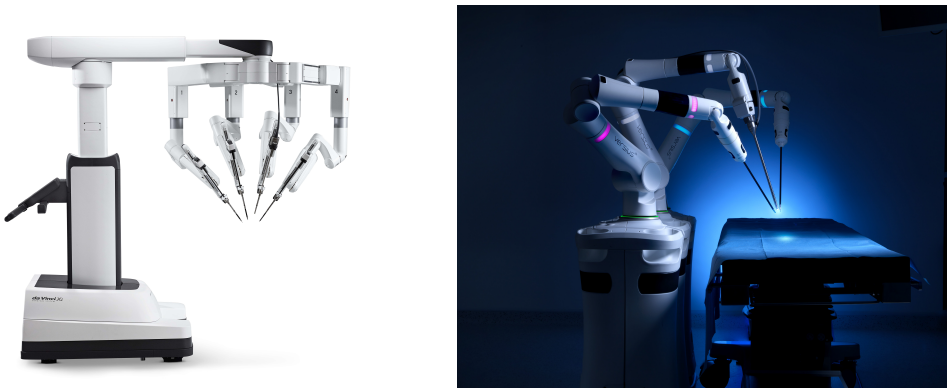
benefits. Multiple studies suggest ergonomic advantages for surgeons [Monfared 2022; Zárate Rodriguez 2019], which prevent operating burnout [Wells 2019] and improve longevity [Stucky 2018]. This evidence can best be understood through the surgery that lead to the initial success of the da Vinci® surgical robot, the minimally invasive prostatectomy. The prostate, being located low (inferiorly) in the pelvis, is hardly accessible via laparoscopic tools. The large distance from trocar to prostate makes precise motion difficult and creates a lever that generates torques which cause fatigue. A robot makes it ergonomically much simpler to operate in these areas. In England, about 85% of radical prostatectomies in 2017 were performed robotically [Maynou 2021].

Interviewing a Surgeon To find further qualitative insights on the advantages of RMIS over MIS, we interviewed Nicholas Raison, a consultant urological surgeon, during a course he organized for robotic partial nephrectomy, hosted by the International Medical Robotics Academy at Guy's Hospital King's College London (KCL). Among the aforementioned lever with associated fatigue reduction, and precision improvements, he named four more advantages of robotic platforms. First, classical laparoscopic tools are straight and do not allow for wrist rotations at the grippers. Robotic tools provide additional degrees of freedom and therefore increased dexterity. This simplifies tasks such as suturing significantly. Second, the control of the tools through a console allows the surgeon to operate from a first-person view

which reduces mental burden. Third, RMIS reduces the learning curve for newly trained surgeons, which results from the enhanced dexterity and the first-person view. Finally, since surgeons are in charge of moving the camera themselves in RMIS, they tend to move it more frequently which provides them with a precise view. In classical MIS, there exists a communication barrier with the camera holder, refer Fig. 1.4. Therefore, in classical MIS, camera holders sometimes show a distant view to compensate for this communication barrier. This last point highlights the potential value of endoscope motion automation, enabling a future fusion of RMIS and MIS techniques.

Future Aspects To best current evidence, robotic surgery platforms currently do not provide improved surgery outcomes, but increase operating cost and time. Viewed differently, they do not worsen the surgery outcomes either. What makes robotic laparoscopy so successful are the surgeon side benefits and its future prospects. These e.g. include the possibility to control the systems remotely, data collection, realistic simulations for training, and potential automation. With this in mind, the research carried out in this thesis aims to address the increased cost and time components, emphasizing on laparoscopic camera motion automation. To get a better understanding of how this could best be achieved, the next section provides an overview of current robot surgery platforms.

1.3.2 Robot Surgery Platforms



(a) The da Vinci® Xi system. Images provided with courtesy of ©2023 Intuitive Surgical Operations, Inc.

(b) The Versius® system. Images provided with courtesy of ©2023 Cambridge Medical Robotics (CMR) Surgical, Ltd.

Figure 1.9: Two examples of currently available commercial robotic laparoscopy systems. It is demonstrated how competition has led to two very different designs. The da Vinci® Xi system in Fig. 1.9a is monolithic with a mechanical RCM. The Versius® system is modular with a programmable RCM. Refers to Section 1.3.2.

Table 1.1: A non-exhaustive list of commercial robotic laparoscopy systems. The table differentiates between monolithic / modular systems and systems with mechanical / programmable RCM. The market competition has lead to a variety of systems with a tendency to stand apart from Intuitive's status-quo approach. Refers to Section 1.3.2.

	Mechanical RCM	Programmable RCM
Monolithic	• da Vinci® (Intuitive) • Maestro® (Moon Surgical)	• hinotori™ (Medicaroid)
Modular	• Hugo™ (Medtronic) • SSi Mantra™ (SS Innovations)	• Versius® (CMR Surgical) • Dexter® (Distalmotion) • Senhance™ (Asensus Surgical)

The da Vinci® robot (Fig. 1.9a) was launched in 1999 by Intuitive Surgical, Inc., which filed multiple patents in the 1990s thereby guarding market access from other companies. Patents in the US generally last for 20 years, which is why as of 2020 most of these critical patents have ran out. Many other companies are now trying to gain some of Intuitive's market share. This competition will ultimately benefit patients, as it might help drive down cost for robotic surgery in the future [Patel 2021], increase accessibility, and lead to innovation in general.

A non-exhaustive overview of current commercial systems is given in Table 1.1. It can be seen that the competition has already brought innovation and a broader variety of systems. Multiple companies are adopting Intuitive's monolithic structure, that is all robotic arms are attached to a single instance. Many others, such as Medtronic and CMR Surgical, are taking a modular approach instead. The systems further differentiate themselves not only by their modularity, but also through their mechanical properties. The da Vinci® robot has a mechanical remote center of motion (RCM), i.e. there are dedicated degrees of freedom (DoF) for achieving a zero translational velocity at the entry point to the patient. The RCM is also sometimes called fulcrum point in the medical context. Having a mechanical RCM provides additional safety, but leads to systems that are hard to re-purpose. This is why recent systems are exploring programmable RCMs, in which a RCM is achieved through control theory. This has implications on safety, as sometimes the RCM may not be achieved, but can potentially provide multi-purpose systems.

1.3.3 Enhancing Current Systems

The previous two sections gave an explanation for the rise of robot assisted laparoscopy in Section 1.3.1, and introduced different commercial systems in Section 1.3.2. We identified surgeon side benefits as well as future prospects as the

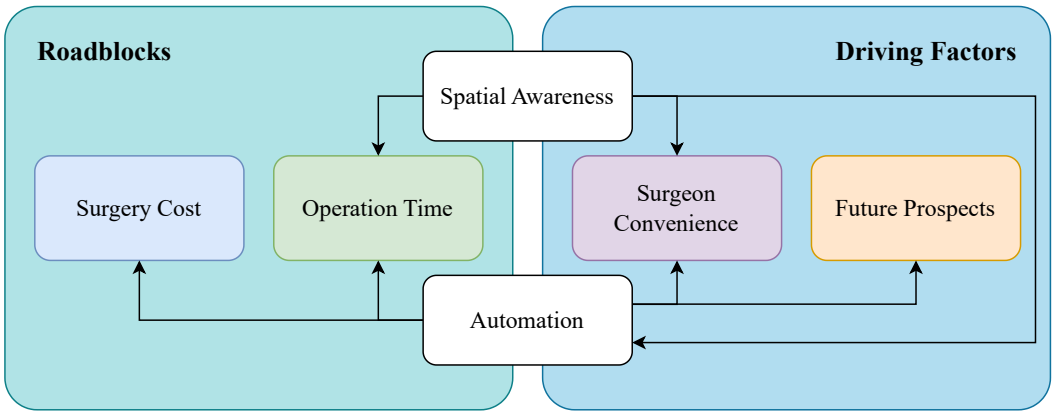


Figure 1.10: Roadblocks and driving factors of RMIS and how the targets of this thesis, spatial awareness and automation, alleviate and enhance them, respectively, refers to Section 1.3.3.

main driving factors for RMIS. Adoption is, however, slowed down by roadblocks such as surgery cost and surgery time. For cholesystectomies, additional costs of about $\$1,000 - 2,000$ per procedure ($\$16,000$ vs $\$18,300$ [Patel 2023]), and additional operation times of about 25 minutes (185 vs 160 minutes [Kane 2020]) are reported. Patient outcomes are just as good in RMIS as they are in MIS. This is to be expected as patient outcomes for MIS are already much better than for open surgery and the bar for further improvement is high. Consequentially, patient outcomes cannot be considered a limitation of current systems and are thus not of immediate relevance for research. Now, to enhance current systems, we argue that it is most efficient to work on the driving factors that made RMIS successful in the first place, as well as the roadblocks. The overarching goal of this thesis, laparoscopic camera motion automation, fits nicely into the clinically relevant enablers and challenges for RMIS, as shown in Fig. 1.10. A prerequisite for automating a robotic system in a pHRI environment, however, is spatial awareness. Spatial awareness comes with the additional benefit that it also contributes to operation time and surgeon convenience, as will be described in the next section.

1.3.3.1 Spatial Awareness

Operating room-level spatial awareness is likely the most underexplored domain in RMIS. By spatial awareness we will be referring to highly precise localization of the robot's links, that is in the order of 1 cm or less, and keep modeling of the rest of the surgical scene for future work.

In the cluttered operating theater, containing an anesthesia machine, monitors, a C-arm, staff, the patient, the surgeon, and many other devices, refer Fig. 1.4, collisions

are inevitable. A staggering amount of 5 arm-arm, and or assistant-arm collisions occur on average in robotic colorectal surgery [Wong 2023]. Pain or bruising from hindrance by the robot arm is reported by 20% of assistants [Vant Hullenaar 2019]. Training assistants could mitigate some of these issues, which, as many studies imply [Cimen 2019; Mitsinikos 2017; Kwon 2020], would lead to improved performance, and, therefore, surgeon convenience and operation time reduction.

To this end, we argue differently and assign the responsibility to the robotic system rather than the assistants or surgeons. These highly capable robotic systems should prevent collisions themselves and leave the clinical workflow unaltered, however, current systems do not model any of the surgical environment and can therefore not generate collision-avoidant actions. Knowing the robot's base frame and infereing location through joint states and forward kinematic could be a first step. Some works suggest image-based avoidance [Hameed 2016] and 3D avoidance [Li 2023a], both of which could benefit from spatial awareness. Collision avoidance might even become more relevant with the adoption of modular systems, as outlined in Section 1.3.2. Spatial awareness would not only improve performance by reducing collisions, it would further contribute to workspace optimization works, such as [Hutzl 2015; Żelechowski 2023], as these require accurate knowledge about the relative patient-robot position. Workspace optimization is not only crucial from a robotic point of view but for clinical aspects as well [Alhusseinawi 2023]. As such, recent studies highlight that it takes an average of 18 minutes to perform proper docking in robotic adrenalectomy [Zuliang 2020].

1.3.3.2 Automation

As already outlined in Section 1.1, laparoscopic camera motion automation is likely the first milestone towards automated RMIS. The short-term benefits of automation are less obvious than for improved spatial awareness, which was explained above, however, the long-term benefits and thus the future prospects justify investigation. Automation, shown in Fig. 1.10, could e.g. have positive implications on surgery cost, operation time, and surgeon convenience.

Some works argue that automation may reduce human error that is linked to fatigue, lack of attention and cognitive overload [Fiorini 2022], therefore contribute to operation time and surgeon convenience. Similarly, automation could help surgeons operate robotic systems by reducing the learning curve [Workum 2018]. On a societal scale, and in an ageing society with shrinking workforce, automation could help to retain accessibility to healthcare and limit cost. This could e.g. be achieved through digital twins of surgeons [Zidane 2023] (CEO of Asensus Surgi-

cal, Senhance™ surgical robot). It is, therefore, expected that parts of RMIS will ultimately be automated [Davenport 2019; Zidane 2023].

In this work, and as will be discussed in the following paragraphs, we will be looking at concrete advantages of automation, like continuous camera motion, as well as strategies that are to be taken into consideration for successful automation of camera motion in laparoscopic surgery.

Automation for Continuous Camera Motion Current RMIS systems, like the da Vinci® or the Hugo™ robots, strictly separate camera and tool motion to prevent collisions. This separation helps to remove some mental burden from the surgeon, but lowers the efficiency and increases the operation time, which in turn has implications on the surgery cost. Having to switch constantly between camera and tool motion also introduces inconvenience to the surgeon. Camera motion automation could help alleviate all these by introducing continuous camera motion. In fact, and as explained in Section 1.3.1 - Interviewing a Surgeon, RMIS already leads to more frequent camera motion when compared to MIS. It is therefore expected that even more camera motion is desirable for surgeons.

Automation in Robot-free Surgeries Surgery cost might not only be reduced through continuous camera motion, it might also be reduced in procedures where robot assistance is currently not common practise, such as laparoscopic cholecystectomy. In Michigan, United States, e.g., only 7.5% of all laparoscopic cholecystectomies were performed with a robot [Sheetz 2020]. The numbers are likely lower in other countries. Domains without robot dominance could initially benefit from replacing the camera holder assistant, refer Fig. 1.4, thus reducing the cost drastically and also allowing the assistant to perform more meaningful tasks. This could further remove the communication barrier between surgeon and camera holder, therefore improving surgeon convenience and reducing the operation time.

Indeed, some companies, e.g. Moon Surgical (Maestro®), CMR Surgical (Versius®), and Distalmotion (Dexter®), target this area through their modular systems, even if they are not introducing automation yet. Due to the progressive nature of healthcare [Chatterjee 2024], it is almost certain that automation will grow in importance as robots will be deployed to broader laparoscopy. Automating camera motion in domains without robot dominance introduces implications that will become clearer in Section 1.5.5. We hold the opinion that any laparoscopic camera motion automation attempt should be compatible with surgeries that are robotic and those that will be robotic or hybrid.

Automation and System Considerations Whilst automation can be achieved in robotic systems with mechanical RCM and systems with programmable RCM, refer Section 1.3.2, in this work, the main focus will be put on systems with programmable RCM. This is inline with our belief that modular systems will increase their market share, as well as the availability of experimental platforms of relevance. Systems with programmable RCM can be used flexibly across multiple types of surgeries, e.g. spine [Farber 2021] or orthopaedic surgery [Suarez-Ahedo 2023], thereby reducing cost even more. To facilitate the potential cost reduction that arise with programmable RCMs, methods in this thesis should therefore not be limited to, but also obey the principles of programmable RCMs.

1.4 Spatial Awareness in Robotic Laparoscopy

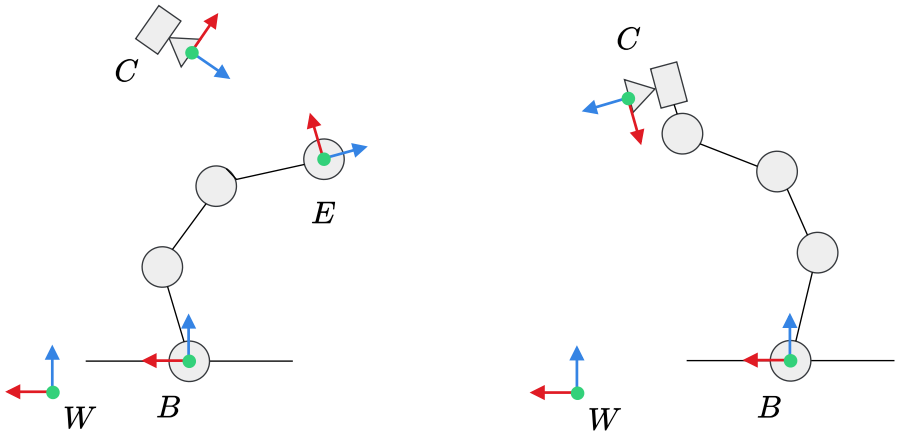
Section 1.3.3 outlined the need for improved spatial awareness in robotic laparoscopy, i.e. knowing the robot's and the laparoscope's locations. It is a prerequisite for automation and facilitates improved clinical workflow for reduced surgery time and surgeon convenience but also adds clinically relevant workspace knowledge. Since we are interested in adding spatial awareness through low cost sensors, i.e. cameras, in this section we will initially summarize camera intrinsic parameter calibration in Section 1.4.1, which is a prerequisite for all that follows. Then, the concepts of eye-in-hand and eye-to-hand calibration are discussed in Section 1.4.2, see Fig. 1.11. Next, we evaluate the clinical workflow implications for these types of calibrations in Section 1.4.3.1, and finally we propose novel methods for marker-free calibration that do not alter the clinical workflow Section 1.4.3.2.

1.4.1 Camera Intrinsic Parameter Calibration

The following section describes camera intrinsic parameter fundamentals and is added for completeness. One may e.g. refer to [Zhang 2023].

Camera intrinsic parameters describe how 3D observed points are projected onto the 2D image plane. This image formation process is crucial for relating 3D points to 2D observations and vice-versa. Let ${}^{W}\mathbf{p}$ be a 3D point expressed in homogeneous coordinates with respect to the world frame W, ${}^{W}\mathbf{p} = \begin{bmatrix} {}^{W}\mathbf{x} & 1 \end{bmatrix}^{T}$, ${}^{W}\mathbf{x} = \begin{bmatrix} {}^{W}x & {}^{W}y & {}^{W}z \end{bmatrix}^{T}$. It is transformed into the camera coordinate frame C via a homogeneous transform ${}^{C}\mathbf{\Theta}_{W}$ as

$$ {}^{C}\mathbf{p} = {}^{C}\mathbf{\Theta}_{W} {}^{W}\mathbf{p}. \tag{1.1} $$



(a) Eye-to-hand setup.

(b) Eye-in-hand setup.

Figure 1.11: Eye-to-hand, and eye-in-hand setups for serial arm manipulators. Camera frames C, robot base frame B, end-effector frame E and world frame W. Real setups might consist of multiple robots. Refers to Section 1.4.

Under the pinhole camera model, the point ${}^{C}\mathbf{p}$ is then projected onto the image plane via the intrinsic camera parameters $\mathbf{K}$

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{1.2}$$

containing the focal lengths $f_{x/y}$, and the principal point $c_{x/y}$. Thus, the observed point in homogeneous image coordinates $\mathbf{u} = \begin{bmatrix} u & v & 1 \end{bmatrix}^T$ is obtained through

$$s\mathbf{u} = \mathbf{K}\mathbf{P}\,{}^{C}\mathbf{\Theta}_{W}\,{}^{W}\mathbf{p}, \tag{1.3}$$

With $\mathbf{P} = \begin{bmatrix} \mathbf{I}^{3\times3} & \mathbf{0}^{3\times1} \end{bmatrix}$, simply being a projection matrix, and s being a scalar usually set to $s = z_C$ when $z_C \neq 0$. Therefore yielding

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f_x x_c / z_c + c_x \\ f_y y_c / z_c + c_y \end{bmatrix}. \tag{1.4}$$

This projection assumes perfect optics, which is not the case in general. Real-world optics introduce radial and some tangential distortion. It is for this reason that

distortion is accounted for via

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f_x x'' + c_x \\ f_y y'' + c_y \end{bmatrix}, \tag{1.5}$$

with

$$\begin{bmatrix} x'' \\ y'' \end{bmatrix} = \begin{bmatrix} x' \frac{1+k_1 r^2+k_2 r^4+k_3 r^6}{1+k_4 r^2+k_5 r^4+k_6 r^6} + 2p_1 x' y' + p_2 \left(r^2 + 2x'^2\right) + s_1 r^2 + s_2 r^4 \\ y' \frac{1+k_1 r^2+k_2 r^4+k_3 r^6}{1+k_4 r^2+k_5 r^4+k_6 r^6} + p_1 \left(r^2 + 2y'^2\right) + 2p_2 x' y' + s_3 r^2 + s_4 r^4 \end{bmatrix}, \tag{1.6}$$

$$r^2 = x'^2 + y'^2, \tag{1.7}$$

and

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x_c / z_c \\ y_c / z_c \end{bmatrix}. \tag{1.8}$$

Therein, k_i are radial distortion coefficients, p_i tangential distortion coefficients, and s_i thin prism coefficients. A camera calibration is performed through observing a pattern of known dimensions, e.g. a checkerboard, from various angles and positions, and optimizing for the model parameters, i.e. camera intrinsic parameters and distortion coefficients, such that the known projected points equal the observed and undistorted points.

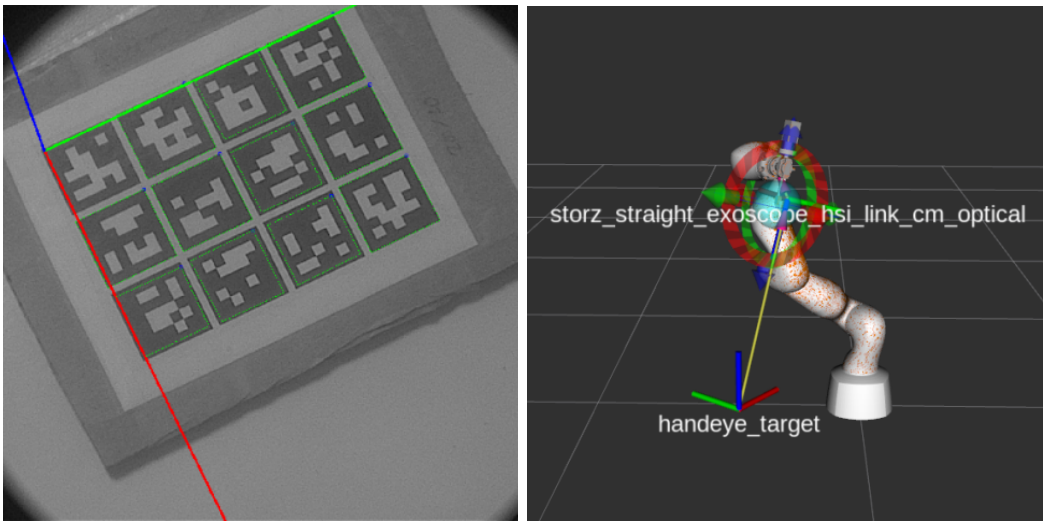
1.4.2 Eye-in-hand and Eye-to-hand Calibration

The following section describes hand-eye calibration fundamentals and is added for completeness. One may e.g. refer to [Ma 2014].

An overview of eye-in-hand and eye-to-hand setups is shown in Fig. 1.11. Eye-in-hand calibration refers to the process of finding the camera frame C to robot end-effector frame E transformation ${}^{E}\boldsymbol{\Theta}_{C}$ when the camera is attached to the robot's end-effector. Eye-to-hand calibration refers to finding the camera frame C to robot base frame B transformation ${}^{B}\boldsymbol{\Theta}_{C}$ when the camera is statically placed next to the robot. A prerequisite for this type of calibration is a well known camera model, as was described in Section 1.4.1.

The calibration is performed by collecting camera to target ${}^{T}\boldsymbol{\Theta}_{C}$ and end-effector to

robot base transformations ${}^{E}\mathbf{\Theta}_{B}$. Given undistorted camera images, and the camera intrinsic parameters, refer Section 1.4.1, camera to target transforms ${}^{T}\mathbf{\Theta}_{C}$ can e.g. be obtained through solving a perspective-n-point (PnP) problem, where points are commonly obtained through ArUco markers, see Fig. 1.12. The base to end-effector transformation ${}^{E}\mathbf{\Theta}_{B}$ can be obtained through robot kinematics. In both,



(a) Undistorted exoscope view of the ArUco marker target.

(b) Mesh visualization of robot state with estimated camera and target poses.

Figure 1.12: Eye-in-hand calibration example in a realistic scenario. Used hardware includes: Storz VITOM Telescope 0° w Integ. Illuminator., Storz TH 102 H3-Z FI camera head, and KUKA LBR Med7 R800. Refers to Section 1.4.2.

the eye-in-hand and eye-to-hand scenario, it is known that the transformation from end-effector to camera ${}^{C}\mathbf{\Theta}_{E}$ and target to end-effector ${}^{E}\mathbf{\Theta}_{T}$, respectively, remains unchanged. Therefore, recording transforms, e.g. for the eye-in-hand setup, in two configurations lets one determine

$$ {}^{B}\mathbf{\Theta}^{1}_{E}\,{}^{E}\mathbf{\Theta}_{C}\,{}^{C}\mathbf{\Theta}^{1}_{T} = {}^{B}\mathbf{\Theta}^{2}_{E}\,{}^{E}\mathbf{\Theta}_{C}\,{}^{C}\mathbf{\Theta}^{2}_{T}, \tag{1.9} $$

which can be re-arranged as

$$ {}^{E}\mathbf{\Theta}^{2}_{B}\,{}^{B}\mathbf{\Theta}^{1}_{E}\,{}^{E}\mathbf{\Theta}_{C} = {}^{E}\mathbf{\Theta}_{C}\,{}^{C}\mathbf{\Theta}^{2}_{T}\,{}^{T}\mathbf{\Theta}^{1}_{C}, \tag{1.10} $$

thus yielding an equation of the form $\mathbf{AX} = \mathbf{XB}$ with $\mathbf{A} = {}^{E}\mathbf{\Theta}^{2}_{B}\,{}^{B}\mathbf{\Theta}^{1}_{E}$ and $\mathbf{B} = {}^{C}\mathbf{\Theta}^{2}_{T}\,{}^{T}\mathbf{\Theta}^{1}_{C}$. Solving for $\mathbf{X} = {}^{E}\mathbf{\Theta}_{C}$ provides the desired eye-in-hand transformation. Similarly, a system of the form $\mathbf{AX} = \mathbf{XB}$ can be derived for the eye-to-hand setup.

A system of this form can e.g. be solved using [Tsai 1989; Park 1994; Horaud 1995].

1.4.3 Unified Calibration for Optimal Clinical Workflow

Having a solid understanding of the calibrations that are required for this thesis, we will now investigate the applicability within the clinical context. When applied within the clinical scenario, e.g. for collision avoidance or autonomous control, as was explained in Section 1.3.3, it is of utmost importance that the clinical workflow must not be disrupted. In this section, implications of the different calibrations to the clinical workflow will be discussed. Following that, the idea of marker-free calibration is introduced, and it is explained how calibration related clinical workflow constraints could be resolved.

1.4.3.1 Implications of Calibrations to the Clinical Workflow

As discussed in the above, Section 1.4.1 and Section 1.4.2, calibrations require calibration targets. In current practise, the necessity of a calibration target would add overhead to the clinical workflow. Calibrations also depend on the type of camera used, which refers back to Section 1.3.3.2 - Automation in Robot-free Surgeries. Standard laparoscopic camera heads are detachable for sterilization of the laparoscope and provide an adjustable focus. Both actions potentially alter the camera intrinsic parameters and make re-calibration necessary. For the remainder of this thesis, we will therefore assume that the focal length stays fixed in a range appropriate for surgery. Furthermore, we will assume that the camera head does not rotate about the optics. This assumption is, however, not as crucial for the camera intrinsic calibration, since surgery optics are of extremely high quality with very little distortion. It is more relevant for eye-in-hand calibrations, where the obtained camera pose is crucial for control. For any external camera, these assumptions do not need to be made and they are allowed to move freely.

The implications of calibrations to the clinical workflow refer back to Fig. 1.10, where additional operation time and surgeon convenience were identified as vital aspects for the shortcomings and advantages of RMIS, respectively. An overview of the implications of the different calibrations is provided in Table 1.2. As can be seen, only the camera intrinsic parameter calibration, in principle, has very little impact on clinical workflow. This, again, is not always the case in classical laparoscopy, especially for the commonly used detachable camera heads, e.g. Fig. 1.12, which, depending on the mounting orientation and adjustable focal length, influence camera intrinsic parameters and eye-in-hand calibration parameters. For automation purposes, it should therefore be enforced that these parameters are known and re-

Table 1.2: The implications of different calibrations to the clinical workflow. Refers to camera intrinsics calibration (Section 1.4.1), eye-in/to-hand calibration (Section 1.4.2), and Section 1.4.3.

Calibration Type	Offline	Clinical Workflow Implications
Camera intrinsics	Cond.	Camera head ideally static with fixed focus
Eye-in-hand	Cond.	Camera pose fixed and known, else re-calibrated
Eye-to-hand	No	Calibration online. Devices may not be moved

main unchanged. For spatial awareness, the constraints are less restricting, since an external camera suffices to estimate robot poses. The external camera's intrinsic parameters can fully be determined offline, therefore keeping the clinical workflow untouched. In this case, it is still necessary to use a calibration target, which introduces inconvenience and mental burden to the surgeon and clinical staff. This reason contributes to surgical robots still being currently spatially unaware. We will therefore introduce the idea of marker-free calibration for eye-to-hand setups, and eye-in-hand setups.

1.4.3.2 Marker-free Unified Calibration

To address the needing of calibration targets, recent works have focused on utilizing the robot itself as calibration target. The idea is simple, the robot mesh, given the current robot configuration, see e.g. Fig. 1.12, is compared against the real robot, thus allowing to estimate camera position with respect to the robot. Initial work estimates joint positions from images and solves a PnP problem, given the robot kinematics to estimate poses [Labbé 2021]. Other work performs differentiable rendering for increased accuracy [Chen 2023], and some turn PnP and differentiable rendering into a self-supervised problem [Lu 2023].

Since we are mostly interested in estimating the robot pose in eye-to-hand scenarios, as was described in Section 1.3.3, and due to the difficulties with eye-in-hand setups with laparoscopic cameras, as was explained in the previous section, this work will focus on using external cameras for pose estimation rather than the laparoscope itself. We will be assuming access to a stereo camera, therefore, simplifying the registration task. Access to stereo images should improve registration reliability, which is important for surgery.

1.5 Camera Motion Automation in Robotic Laparoscopy

Having established a proper understanding of the importance of spatial awareness for RMIS in general (Section 1.4), and for automation in particular, this section will investigate camera motion automation in robotic laparoscopy, which was identified as the second desired enhancement in RMIS, Section 1.3.3, Fig. 1.10. Initially, several alternatives for laparoscopic camera motion automation will be discussed in Section 1.5.1. Of these methods, and due its promising properties, visual servoing will be investigated further. Starting with rule-based approaches to visual servoing and relevant auxiliary tasks like tool segmentation in Section 1.5.2, the reader will be introduced to the limitations of such methods, that is assumptions to laparoscopy that generally do not hold in real surgery. To find a potential solution, the focus is shifted towards data-driven methods. The data-driven methods include RL and IL. Since data-driven methods are relatively new to laparoscopic camera motion automation, a broad review of RL will be given in Section 1.5.4, followed by IL in Section 1.5.5.

1.5.1 Camera Motion Automation Approaches

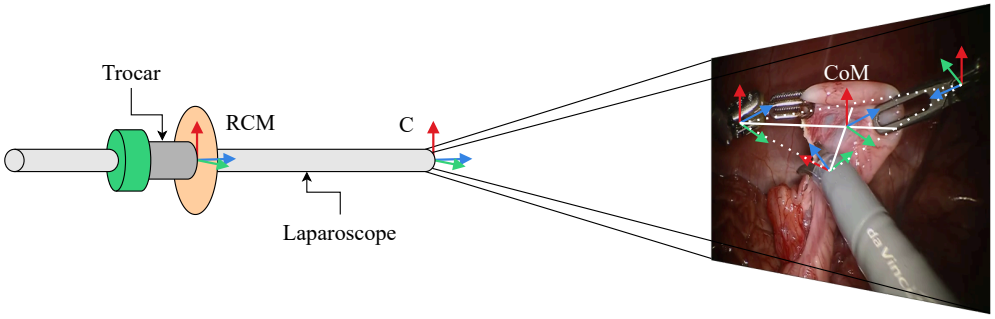


Figure 1.13: Coordinate frames relevant for laparoscopic camera motion automation. Camera frame C, CoM frame CoM, and RCM at trocar. The camera frame C is commonly obtained via eye-in-hand calibration, Section 1.4.2 and Fig. 1.12. For visual servoing, the CoM assumption as view center-point is commonly made. Laparoscopic view shows an image from a da Vinci® system in the SurgVisDom[Zia 2021] dataset. Refers to Section 1.5.2.

Laparoscopic camera motion can be automated in a plenitude of ways. A typical laparoscopic setup is shown in Fig. 1.13. Therein, the ultimate goal is to control the pose of the camera frame C under the RCM constraint. The camera frame C, as was explained in Section 1.4.1, and Section 1.4.2, Fig. 1.12, can be obtained through camera intrinsic parameters plus eye-in-hand calibration.

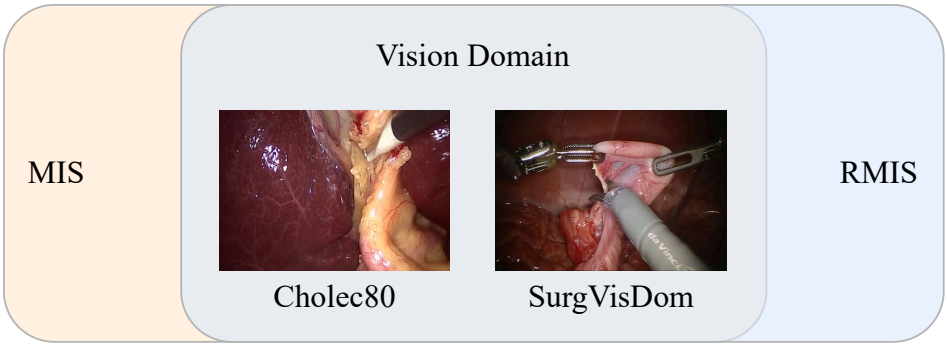


Figure 1.14: The vision domain is a shared domain between MIS and RMIS procedures. Laparoscopic views taken from Cholec80 [Twinanda 2017], and SurgVisDom [Zia 2021]. Refers to Section 1.5.1.

In fully robotic setups, a common approach to automation is to simply use the kinematic data that is available through the joint position encoders [Col 2020]. The camera then follows some geometric point, like the CoM between the tools, as shown in Fig. 1.13. Without the kinematic data, these methods are not applicable to hybrid setups, which are of particular interest to this work, as was mentioned in Section 1.3.3.2 - Robot-free surgeries. They further suffer some other shortfalls, such as the inability to obey anatomic constraints and reliance on accurate multi-arm calibrations. Other, semi-autonomous techniques, aim to alleviate some of the model uncertainties by assigning the surgeons a greater responsibility to controlling the camera in a collaborative fashion. Gaze or voice control are among them [Taniguchi 2010]. They, however, lead to eye strain, additional mental workload or communication failures, and do not satisfy the - level five: high autonomy - target that was set in Section 1.1.

Alongside automation via kinematic data, visual servo work we present a very simplistic registration proview through an image space loss, is considered a promising alternative, as it provides intra-operative feedback [Pandya 2014] and is less prone to errors from model mismatch [Azizian 2014]. It is capable of understanding and interpreting the surgical scene, thus potentially enabling level five autonomy and above, the overarching goal of this work, Section 1.1. Visual servoing, in itself, is of special interest to this work for one additional reason. It characterizes MIS to RMIS transferability, as was targeted in Section 1.3.3. Vision poses a shared domain between MIS and RMIS, the vision domain, as shown in Fig. 1.14. Visual servoing methodologies could thus be transferred from MIS to RMIS and vice-versa. In the next section, we will, therefore, review rule-based visual servoing approaches.

1.5.2 Rule-based Visual Servoing

Rule-based visual servoing approaches are a well established research field for laparoscopic camera motion automation. These methods are formulating control through specifying some proxy for autonomy. There exists research on visual servoing with mechanical RCM and visual servoing with programmable RCM. Although both areas aim at controlling the camera frame, refer Fig. 1.13, we will distinguish between them for better clarity and to comply with the targets that were outlined in Section 1.3.3. The following two paragraphs will, therefore, review current rule-based approaches with mechanical and programmable RCM, respectively, followed by a paragraph that analyzes the shortcomings of these methods.

Visual Servoing with mechanical RCM Approaches that use a mechanical RCM are [Omote 1999], where a visual servo is implemented to control the CoM of a colored marker on a forceps in image space. In [Agustinos 2014; Voros 2007], the tool tip position is found in image space via kinematic knowledge over the tool entry points and a visual servo is applied to center the tool tip in image space. Another common scheme is to alter the camera's zoom based on the distance of the surgical tools, which was first presented in [King 2013], where the authors use colored markers to track the surgical instruments.

The authors in [Eslamian 2020; Mariani 2020; Col 2020], with related prior works in [Eslamian 2016; Eslamian 2017], compute the center point in between two surgical tools via their respective positions and align the camera's optical axis with the line that spans from RCM to the tools' center point, which requires a complicated registration procedure. Hongwei Li et al. [2016] also rely on the positions of the surgical tools and adjust the field of view's width based on the distance of them. Abdelaal et al. [2020] use a similar approach as Eslamian et al. [2020], in that they adjust the camera distance to the surgical scene based on the tool distance, however, they do not align the camera's optical axis with the line that spans from RCM to the camera, but rather with the scene's surface normal, which is made possible by their 6 DoF endoscope. In [Ma 2019], X. Ma et al. deploy a visual servo to center a green marker on a tool by incorporating depth information as extracted from camera and tool motion. In [Ma 2020], they extend this work into a quadratic program in which they minimize joint velocities whilst constraining the camera's distance with respect to the tools and the average tool position in the image plane to be central, where they rely on stereoscopic images to extract depth information. Gruijthuijsen et al. [2022] propose a framework for semantically rich collaborative control but effectively only track surgical tools.

Visual Servoing with programmable RCM Where a mechanical RCM is not available, it can be achieved programmatically. As such, Aghakhani et al. [2013] design a composite Jacobian method that integrates a RCM objective with a task function that defines an error on points in image space. The authors in [Yang 2019] also design a Jacobian gain controller that enforces the tip of a tool to reside within a defined region by computing the winding number of that region around the desired point. They additionally request the endoscope to extend the surgeon's natural line of sight. In [Li 2020a], W. Li et al. introduce the RCM and a visual error via the image Jacobian as constraints to a quadratic problem that aims at satisfying these constraints whilst minimizing the joint velocities. Sandoval et al. [2021] propose a torque control framework that includes remote center of motion constraints, tool center point following and nullspace projects for arm-staff collisions.

Flaws of current Visual Servos It becomes apparent that most of these methods rely on the mere tool distance to infer a control law, whereas only in [Ma 2019; Ma 2020; Aghakhani 2013; Yang 2019; Li 2020a] the image points for visual servoing can be chosen arbitrarily. This leaves most of the current methods with some fundamental flaws. First, the assumption that laparoscopic camera motion only originates from tool motion, but not from surrounding tissue or organs. However, clinical evidence suggests camera motion is also caused by the surgeon's desire to observe tissue [Ellis 2016]. Surgeons might be interested in examining specific anatomy, e.g. for establishing the critical view of safety, refer Section 1.2.2.2. This claim can easily be verified by the interested reader through watching videos of laparoscopic interventions. Second, all of these methods are of reactive nature and none of them anticipates future potential views. Only in [Weede 2011] and [Ji 2018], the authors consider predictive models based on expert demonstrations. In [Weede 2011], Weede et al. cluster gripper positions in observed interventions and compute transition probabilities from cluster to cluster by modeling the system as a Markov chain. They predict the probability of future tool positions and enforce the camera's optical axis to point to the future probability weighted sum of the tools' positions, which relies on kinematic information. Ji et al. propose in [Ji 2018] to rank image features in object bounding boxes according to expert demonstrations with a linear regression, however, they do not consider an RCM or any constraints in motion that arise from it and only regress their model on a synthetic environment. It remains questionable whether this method could be transitioned to a real setup.

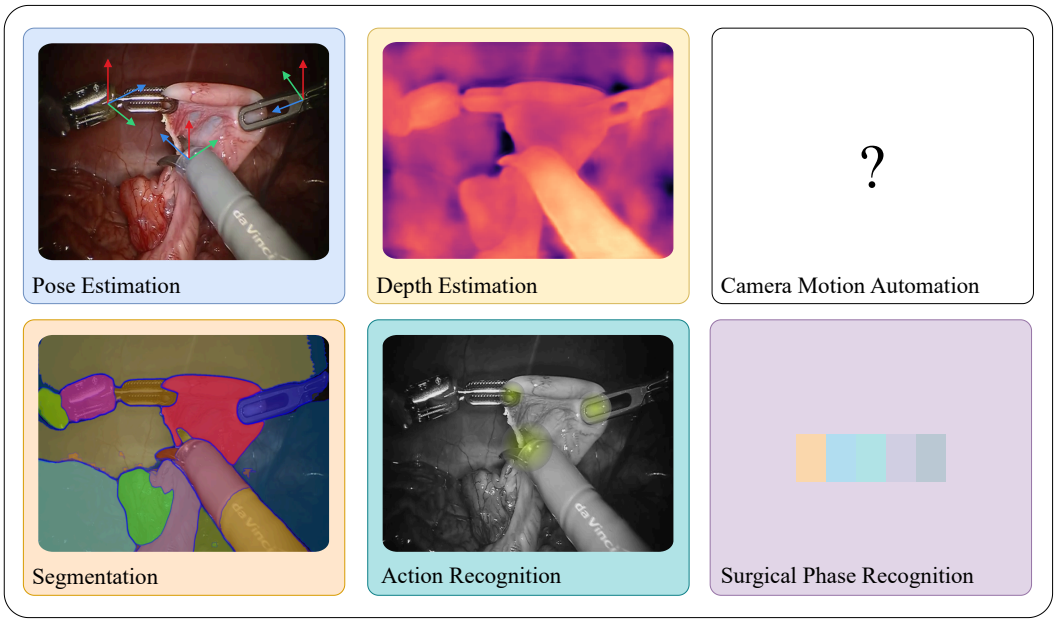


Figure 1.15: Auxiliary tasks that could be used for laparoscopic camera automation but that are not used in practise. None of the data-driven methods directly attempts laparoscopic camera motion automation in realistic scenarios. The surgical phases refer e.g. back to Section 1.2.2. Laparoscopic view shows an image from a da Vinci® system in the SurgVisDom [Zia 2021] dataset. Monocular depth estimated using [Oquab 2024], segmentations generated through [Kirillov 2023]. Refers to Section 1.5.3.

1.5.3 Auxiliary Vision Tasks

Apparently, current visual servos oversimplify the automation task greatly. Only in [Gruijthuijsen 2022], Gruijthuijsen et al. incorporate neural networks for autonomous instrument tracking. What is generally lacking is an understanding of the surgical scene that could help overcome the simplifications that are currently made for visual servoing. Indeed, there exists plenty of research on solving auxiliary vision tasks through data-driven methods, see Fig. 1.15. Some of which, will be highlighted below.

Automation-related Tasks Based on progress in segmentation tasks, using deep learning, but also with the prospect of identifying tool tips for automating procedures, a vast body of literature on surgical tool segmentation evolved, including [Allan 2019; Pakhomov 2019], and [García-Peraza-Herrera 2017b; García-Peraza-Herrera 2017a; Shvets 2018; Islam 2019; Jin 2018a; Sarikaya 2017; Costa Rocha 2019; Attia 2017; Laina 2017; Garcia-Peraza-Herrera 2021]. Some of these segmentation works help improve surgical tool pose estimation, as was shown in [Kur-

mann 2017; Li 2020b; Hasan 2021], as does monocular depth estimation [Li 2021b; Huang 2021; Xu 2022; Shao 2022; Li 2023b; Lou 2024; Budd 2024] and [Li 2022c; Bardozzo 2022; Huang 2022; Huang 2021]. Whilst these works could be fused with any of the visual servos from above, it wouldn't resolve the general assumption that camera motion only results from tool motion. Besides tool segmentation, research exists for surgical phase detection [Stauder 2014; Lalys 2014; Dergachyova 2016; Twinanda 2017; Malpani 2016; Jin 2018b; Ross 2018; Yengera 2018; Funke 2018; Yu 2020] as well as [Bodenstedt 2019a; Padoy 2019; Czempiel 2020; Jin 2020; Kitaguchi 2020]. These could condition visual servos on the current phase of the surgery. Other self-supervised approaches, which could be integrated similarly, aim to estimate the remaining time of a surgical procedure [Twinanda 2019; Bodenstedt 2019b; Rivoir 2019] or perform action recognition [Nwoye 2021].

The Automation Task Automation-related tasks, as introduced in the previous paragraph, are often treated as prerequisite for autonomy, but could at present only contribute as input to a smarter visual servoing scheme. Given the success of data-driven methods for these auxiliary tasks, it might be reasonable to utilize them for laparoscopic camera motion automation as well. But instead of taking any of the auxiliary tasks as priors, one might consider learning automation directly, too. This is because, firstly, in deep learning, end-to-end approaches have proven to outperform methods that rely on hand-crafted inputs that may seem humanly logical, and, secondly, it is the simplest approach. So instead of extracting redundant information, such as tool tips, surgery phases, depth, or poses first, we suggest that it might make sense to apply data-driven methods to automating laparoscopic camera motion immediately, whilst keeping the robot through classical control in charge of executing the learned camera motion. This combination of learning and classical control could facilitate safe execution of complex behaviors beyond the tool following proxy. Interestingly, data-driven laparoscopic camera motion automation is an underexplored field. In spite of the underexploration of data-driven methods for laparosopic camera motion automation, we turn to a broader body of literature, and introduce several learning-based methods in the next sections, namely RL in Section 1.5.4, and IL in Section 1.5.5.

1.5.4 Reinforcement Learning

RL is conceptually simple as it draws inspiration from human or animal trial-and-error learning. In a RL scenario, see Fig. 1.16, an agent interacts with an environment. The agent observes the state s_t and performs an action based on it. The agent will then observe the changed state s_{t+1} and receive a reward r_{t+1}. Through

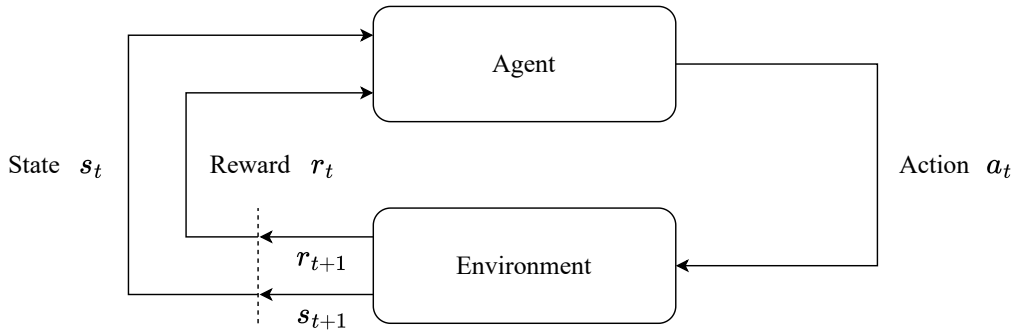
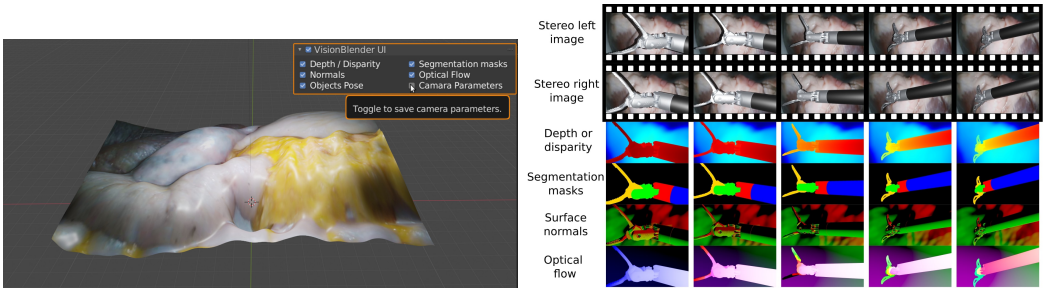


Figure 1.16: A procedural diagram of RL. Given the environment state s_t, an agent performs an action a_t and observes the resulting state s_{t+1} and reward r_{t+1}. Refers to Section 1.5.4.

trial and error, the agent will try to maximize the reward. RL is sample inefficient and requires a large amount of trials until the agent learns to solve a task successfully. RL is in fact so sample inefficient, that it is often necessary to first train in simulation, where physical systems can be mimicked far beyond realtime. If a system can be simulated well and reward functions can be specified, then RL can find impressively complex behaviors in vast state spaces that could not be explored through classical search algorithms. This is why there exists impressive research in RL, where it was e.g. shown that agents can learn to play Atari games just through observing images [Mnih 2013]. This success lead to systems that are capable of beating humans in the game of Go [Silver 2016], which was later improved to learn entirely through self-play [Silver 2017]. For reference, there exist many more possible states in Go than there are atoms in the known universe. These two examples stem from relatively simple simulation environments (Go and Atari games). But even in physically more challenging scenarios, it was shown that humanoid walking can be solved through RL [Schulman 2017]. The transferability from simulation to the real systems, however, was left unsolved. The most recent advancements succeeded in transferring highly complex policies from simulation to real robots, like standing up [Rudin 2022] or parkour [Hoeller 2023] on the ANYmal quadruped, and playing football on a simplified humanoid robot [Liu 2022]. The state of simulation in surgery will be summarized in the next paragraph.

Laparoscopic Camera Motion Automation The hurdles for RL in laparoscopic camera motion automation are obvious. The sample inefficiency and potential harm to patients currenty restrict RL approaches to simulation [Su 2021; Agrawal 2018]. Significant steps are being made to improve simulations, like [Scheikl 2023] using e.g. SOFA [Allard 2007], but a domain gap remains. Research that at-



(a) Static surgical 3D scene. (b) Generated masks.

Figure 1.17: Blender plugin, named VisionBlender, for rendering realistic surgical scenes. Images with courtesy of [Cartucho 2021]. Refers to Section 1.5.4.

tempts to bridge the domain gap to make RL algorithms deployable in real setups exists, [Cartucho 2021] e.g. provide a Blender plugin for realistic view generation, see Fig. 1.17, and [Marzullo 2021] propose image domain transfer, but only static scenes are considered for now. Clinical translation using RL has yet to be achieved.

We conclude that RL does not comply with the objective of this thesis, that is near term automation that benefits the patient, refer Section 1.1. This is because simulations have not yet reached the realism that is required for narrowing the domain gap to surgery. Therefore, the next Section 1.5.5, will investigate other data-driven methods instead.

1.5.5 Imitation Learning

The goal of IL is to copy the behavior of an expert demonstrator. Hence, IL aims to extract an expert policy $\pi_{\mathrm{E}} : s_t \rightarrow a_t$ that maps a state s_t to an action a_t, where the actions and states are drawn from a trajectory $\tau_i = \{s_t, a_t ..., x_{t+T+1}, a_{t+T+1}\}$, sampled from expert demonstrations $\tau_i \in \mathcal{D}$. The underlying state s_t might not always be fully observable, in which case one observes $\hat{s}_t = f(s_t)$, where f is the unknown mapping of the underlying state s_t to the observed state $\hat{s}_t$. IL without access to the underlying state is often referred to as imitation from observation (IfO) [Liu 2018]. Also, the embodiment of demonstrator and learner might not always be the same, e.g. if the demonstrator is a human and the learner is a robot, which is called a domain shift. IL is usually achieved via either of two dominant approaches [Osa 2018]: behavioral cloning (BC) [Pomerleau 1991] and inverse reinforcement learning (IRL) [Ng 2000]. BC regresses the expert policy π_{E} from sampled trajectories τ_t in a supervised fashion. IRL aims to recover the expert's hidden reward r_t to later optimize a policy to also achieve the recovered reward. Neural networks have become the dominant approach for estimating the

policy π, henceforth, the following sections will focus on them.

1.5.5.1 Behavioral Cloning

In a common BC setup that satisfies the Markov property one takes a current state and tries to predict future actions conditioned only on the current state. One can also condition actions on a set of past states [Xu 2017] but this is rather uncommon. In [Torabi 2018], Torabi et al. try to perform IfO by randomly exploring the action-state space. They learn a mapping from observations to actions and perform BC on newly obtained observations via this mapping. A general issue in BC is covariate shift, that is the inability to generalize from a small dataset. In [Ho 2016], Ho et al. address this issue by introducing generative adversarial learning which implicitly regularizes the policy to a bigger action-state space. Torabi et al. [2019] extend [Ho 2016] by working without immediate access to actions but from observation only. Although generative adversarial IL helps to generalize, a lot of the existing literature focusses on learning an underlying forward dynamics model to infer any policy once the forward dynamics model is known. As such, the authors in [Finn 2016; Finn 2017a; Nair 2017] learn to predict future observations from actions via a dataset of randomly explored action-state space trajectories. They use this state transition model to infer actions that lead to desired states, which requires the user to define a desired state. Other attempts to learn arbitrary behaviors are one-shot and zero-shot IL approaches. In one-shot [Finn 2017b], Finn et al. use model-agnostic meta-learning (MAML) to learn how to learn new tasks quickly. Once the network parameters are initialized via meta-learning, a few gradient steps from a demonstration allow to imitate that demonstration. In [Yu 2018], Yu et al. extend this work across domain gaps, that is a robotic learner imitates a human demonstrator from a single demonstration only. Pathak et al. then introduce zero-shot learning in [Pathak 2018], which introduces a goal conditioned policy, that allows to immediately execute a policy from intermediate pre-defined goal states. In [Hausman 2017], Hausman et al. extend the work in [Ho 2016] by conditioning the policy on the intention, which similarly to zero-shot IL, allows to reach intermediate goals. The idea of learning a forward dynamics model is then extended into a compressed feature space by Srinivas et al. in Universal Planning Networks [Srinivas 2018]. Their work conditions latent-space dynamics on actions from demonstrations and they define an optimization framework that finds actions conditioned on a goal state. Most recent approaches, such as [Lynch 2019], learn a latent state representation that categorically clusters different policies as to create interpretable behaviors.

1.5.5.2 Inverse Reinforcement Learning

In IRL the aim is to extract a reward function from a set of expert demonstrations $\mathcal{D}$. In practise this can be achieved by embedding observations $\hat{s}_t$ into a meaningful feature space and by enforcing that a newly obtained policy π and an expert policy π_{E} follow similar trajectory embeddings.

Different methods for embedding exist. A triplet loss can be formulated to pull similar images closer to each other while repelling them from different ones. In [Wang 2014; Schroff 2015], the authors achieve a triplet loss via a simple distance metric while X. Wang et al. formulate it as the angle between features [Wang 2015]. Another option is to use prediction as a proxy for meaningful embeddings [Vondrick 2016; Sermanet 2017; Srivastava 2016; Mathieu 2016], self-supervised clustering [Caron 2018], or, most recently, contrastive learning methods [Khosla 2021], which similarly to [Wang 2015] aim to align features of samples with similar properties.

Sermanet et al. in [Sermanet 2018] used for example a triplet loss on multi-view videos to learn a view-point invariant embedding that can be used to have a learner learn an expert demonstration. Aytar et al. [2018] had an agent learn to reach checkpoints by sampling checkpoints from demonstrations on YouTube and by defining a reward on the alignment between the current state and the desired checkpoint. In few-shot, Xie et al. [2018] learn initial parameters for a network via MAML, a form of meta-learning, to infer goals in demonstrations from a single gradient-step. They then perform RL to replicate a policy that yields these goals. The idea of Universal Planning Networks by Srinivas et al. in [Srinivas 2018] is further extended by Yu et al. in [Yu 2019], which learns a goal metric for RL in an unsupervised manner.

1.6 Imitation Learning for Laparoscopic Camera Motion Automation

In the previous sections, several considerations that are relevant to achieving laparoscopic camera motion automation, including economical, clinical, and technical aspects, were introduced. Plenty arguments went into the decision making and some of which may have already been forgotten. Therefore, in this section, we will revisit the key-concepts in Section 1.6.1. Next, and given these critical considerations, we will hypothesize a method for laparoscopic camera motion automation in Section 1.6.2, followed by several sections on realizing the hypothesized approach.

1.6.1 Revisiting Key Concepts

The Case for Camera Motion Automation In the Foreword, Section 1.1, it was argued that the progressive nature of surgery will ultimately alleviate clinical staff of unfulfilling tasks through achieving level five autonomy, likely first for laparoscopic camera motion. We took the clinicians' perspective in the cholecystectomy procedure, the most common laparoscopic procedure, to investigate automation claims further. We found that the camera holder assistant, see Fig. 1.4, performs a relatively simple and unfulfilling task. This stays in contrast with the importance of the camera holder's role of establishing and maintaining a good view for highlighting critical anatomies to surgeons, Section 1.2.2. We then analyzed the rise of RMIS in Section 1.3.1, and found first indicators that, in RMIS, the camera is moved more frequently when compared to MIS, suggesting that even more camera motion might be desirable. This stays is in line with the initial rise of RMIS. We thus concluded that camera motion automation might be beneficial for the surgeon, the patient, and the clinical staff in Section 1.3.3, Fig. 1.10.

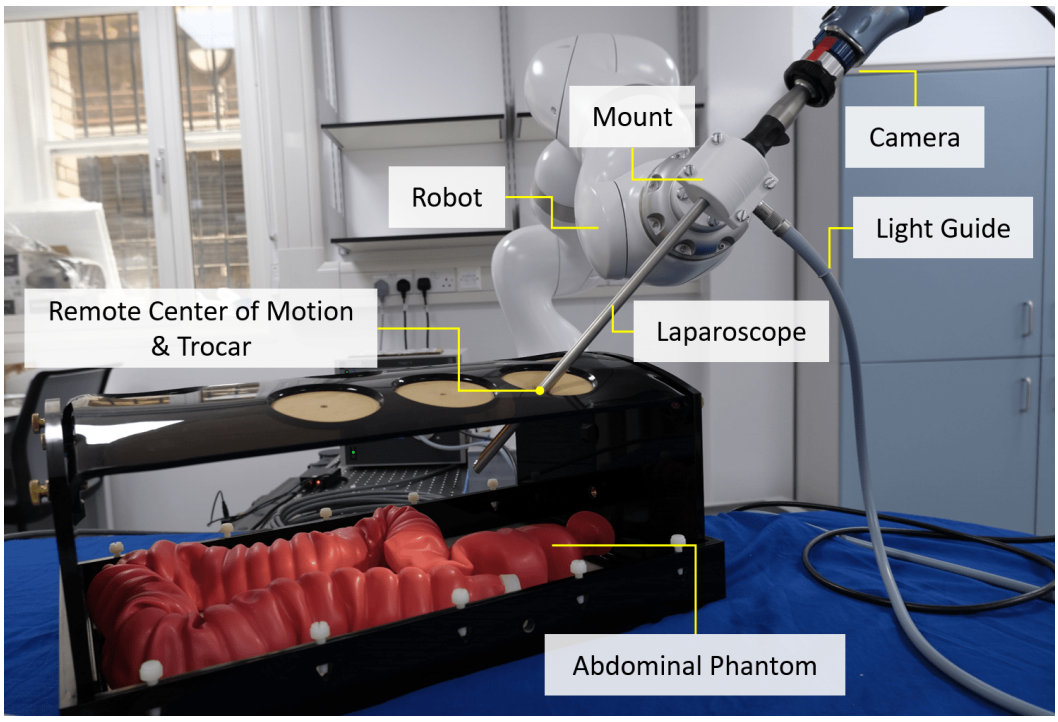


Figure 1.18: Robotic setup. A Storz Endocameleon Hopkins Telescope, which provides a light source port and a camera attachment point, is mounted to a KUKA LBR Med 7 R800 robot via a 3D printed clamp. The robotic system undergoes image-based control to reach desired views of the surgical scene and simultaneously pivots around a programmable RCM.

High-level Aspects of Camera Motion Automation We took economical considerations into account for propagating the clinically beneficial automation changes to the patients with minimal impact on cost. This resulted in the goal of delivering the automation endeavor through a system with programmable RCM to surgeries that are currently robot-free, Section 1.3.3.2. We thus propose a system design that will serve as foundation to this thesis, see Fig. 1.18.

Following the clinical prospects of automation and the economically guided form-factor, we analyzed technical factors. We explained calibration prerequisites, Fig. 1.12, Fig. 1.13, and derived novel means of registration for an improved clinical workflow in Section 1.4.3, which finally lead to automation itself. We discussed several automation approaches in Section 1.5.1, and concluded that learning complex policies, beyond tool following, would require data-driven methods. Crucially, among other reasons, we chose vision as a candidate domain, and thus visual servoing as a candidate for automation, since vision was identified as a shared domain between RMIS and MIS, Fig. 1.14. We found, however, that vision is currently only used for solving auxiliary tasks, Fig. 1.15. We evaluated the state of RL in Section 1.5.4, and concluded that it does not align with our goals of near term level five autonomy, and were thus left with IL approaches in Section 1.5.5. With the robot-free surgery target in mind, and the suggested IL as substrate for automation, we impose embodiment-invariance onto the solution. That is, the expert demonstrator can be human, and the executing agent is a robot. Precisely speaking, we are, therefore, trying to solve IfO, with only access to the partially observed environment state $\hat{s}_t$, see Fig. 1.19. The next sections will go into detail on how this could be achieved concretely.

1.6.2 Hypothesizing Embodiment-invariant Laparoscopic Camera Motion Automation

Having outlined the scope for laparoscopic camera motion automation - embodiment-invariant IfO - further referred to simply as IL, this section will now hypothesize concrete means of achieving it. It will dissect the thought process for framing the path towards autonomy in this thesis rigorously.

IL requires large amounts of data for learning the tail-end, i.e. rare cases. Therefore, we will initially search for available data in Section 1.6.2.1. Given the data, a mixture of supervised and self-supervised methods for learning to predict camera motion, and predicting can be considered imitating, will be suggested in Section 1.6.2.2. Next, a camera motion formulation for the suggested supervised and self-supervised frameworks will be proposed in Section 1.6.2.3, keeping the con-

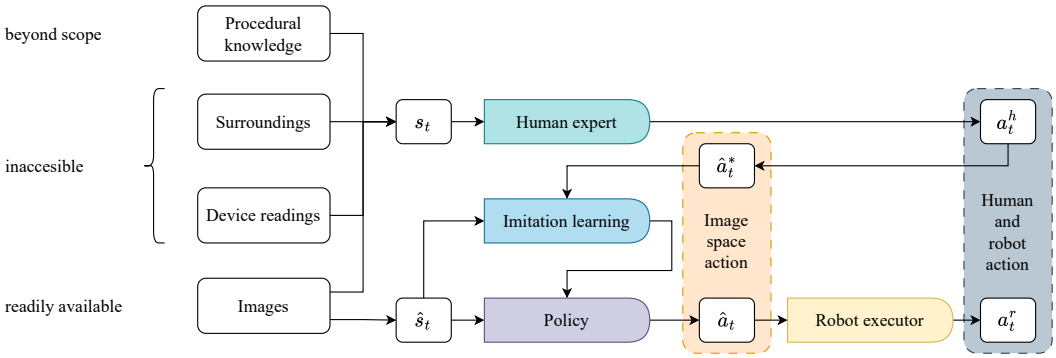


Figure 1.19: The hypothesized approach for laparoscopic camera motion IL. Actions are learned in the shared vision domain (orange) and executed via different embodiments, the human or the robot. The human expert has access to the full environment state s_t and performs an action a_t^h. This action a_t^h, leads to an action in image space $\hat{a}_t^*$, the desired action. We suggest extracting action $\hat{a}_t^*$ from image space for IL purposes, enabling to learn a policy $\pi : \hat{s}_t \rightarrow \hat{a}_t$ that maps the partially observed states $\hat{s}_t$, i.e. images, to actions. The robot executes the predicted action $\hat{a}_t$ in the form of a_t^r via optimal control. Refers to Section 1.6.2.

straint of embodiment-invariance in mind. Finally, Section 1.6.2.4 will explain classical and clinically applicable methods for controlling a robotic laparoscope under the camera motion prediction. An overview of the proposed approach is already shown in Fig. 1.19.

1.6.2.1 Search for available Data

Given the advancements in deep learning and especially in IL, it is surprising that no one applied IL to automate camera motion in laparoscopic surgery. It appears trivial that camera motion could be learned from data of real surgeries, thereby implicitly tackling the domain-gap that e.g. RL methods face. After having a closer look, it comes, however, at no surprise researchers have not tried. The challenge is to collect sufficient data of high quality. In fact, there exists no dataset with state-action (image-camera motion) pairs. Many works agree and highlight that this lack of expert annotated data hinders progress towards camera motion automation in RMIS [Maier-Hein 2022; Kassahun 2016; Esteva 2019].

Data collection is expensive, especially in realistic setups. Recent efforts to make vast amounts of laparoscopic intervention videos publicly available [Maier-Hein 2022] drastically change how IL for camera motion automation could be approached. An overview of the, by the time of this writing, available datasets is shown in Table 1.3. The two biggest datasets, by orders of magnitude, are Cholec80 [Twinanda 2017] and ROBUST-MIS [Maier-Hein 2021], often referred to as HeiChole, both

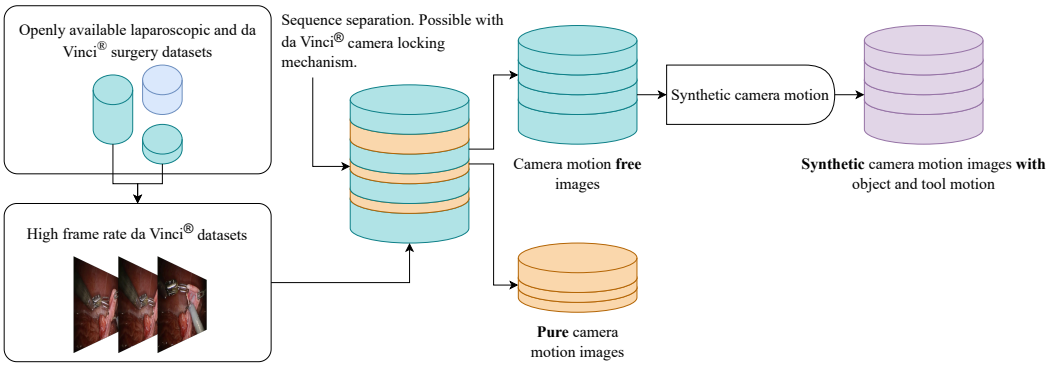


Figure 1.20: The proposed isolation of camera motion, i.e. actions, and tool as well as object motion. The camera locking mechanism of the da Vinci® robot allows for extraction of camera motion free image sequences, see Table 1.3. Synthetically added camera motion can be used for supervised training. Refers to Section 1.6.2.2.

are laparoscopic, which is exactly what we are looking for. Some of the datasets, since they come with hand-annotated segmentations, come at low frame rates, i.e. $1 - 2\,\text{Hz}$, and are, therefore, unusable for IL. As was already pointed out, none of the datasets come with state-action pairs. The only two publicly available datasets that come with kinematic labels are MISAW [Mitsuishi 2013] and JIGSAW [Ahmidi 2017], but both datasets are captured in synthetic environments and without camera motion.

The missing state-action pairs for all of the available datasets are the major roadblock for applying IL methods to them. Reliably extracting camera motion in retrospect from dynamic surgical scenery is an unsolved task in itself. Somewhat surprisingly, one of the da Vinci® datasets, SurgVisDom [Zia 2021], which was intentionally released for domain transfer, plays an important role in extracting camera motion from laparoscopic videos. How we propose to extract laparoscopic camera motion reliably and how the locking mechanism of the da Vinci® robot, refer Section 1.3.3.2, plays a crucial role, will be explained in the following Section 1.6.2.2.

1.6.2.2 Supervised Camera Motion Extraction and Self-supervised Camera Motion Prediction

Paradoxically, camera motion is accessible and intrinsic to videos of laparoscopic interventions, the crux lies in extracting it reliably and isolating it from tool and object motion, for which, currently, no method exists. It is thus not surprising that existing literature on IL for camera motion automation still utilizes data from mock setups [Ji 2018; Wagner 2021]. As was discussed in Section 1.5.3, so far the existing data is mainly leveraged to solve auxiliary tasks that could at best contribute to

camera motion automation. In this section, we will propose a clever trick to extract camera motion from laparoscopic interventions, refer Section 1.6.2.1, as a means of generating state-action pairs.

Extracting Camera Motion from Laparoscopic Interventions The state of the system $\hat{s}_t$, although underdetermined with only access to the images, see Fig. 1.19, is already known. The remaining difficulty is to extract actions from the change between subsequent states, that is extracting camera motion $\hat{a}_t$ from the subsequent states $\hat{s}_t$ and $\hat{s}_{t+1}$. As a reminder, actions lead to a change in the state, nicely shown in Fig. 1.16. It is difficult because it is hard to differentiate between camera, tool, and object motion. Although their might be different methods for achieving this, in this work we propose to train a neural network in a supervised fashion on isolating camera from object and tool motion. A visualization for how we are planning to achieve this is shown in Fig. 1.20. In summary, we propose to use all existing data from da Vinci® surgeries with HFR, see Table 1.3. Next, and as hinted in Section 1.6.2.1, we manually isolate sequences with camera motion from sequences without camera motion. This is made possible by the locking system of da Vinci® robots. Keeping in mind that the sequences without camera motion still contain object and tool motion, we suggest to add camera motion synthetically. The synthetically added camera motion can then be used as ground truth for a neural network. We will be hypothesizing that the learned camera motion extraction will function on videos of laparoscopic interventions to generate image-action pairs.

Camera Motion Prediction Assuming that a sufficiently accurate camera motion extractor can be provided through the above proposal, camera motion prediction should follow swiftly. One would simply take the extracted camera motion as a pseudo-label. This would lead to predicting image space actions $\hat{a}_t$, which are invariant of the domain, refer Fig. 1.19, which was already proposed in Fig. 1.14. The important part is that the extracted as well as the predicted camera motion should be executable on a robotic laparoscope holder. Appropriate means of describing the camera motion will be introduced in the next Section 1.6.2.3.

1.6.2.3 Homography-based Camera Motion Formulation

Extracting camera motion from images is a well studied problem, but not so much in the surgical context, where an extremely dynamic and deformable environment with little texture hardens the task. Therefore, simplifying the problem is crucial. In its simplest form, one can model a surgical scene as a plane, the change in observed views, i.e. induced through camera motion, can then be described via homographies.

Whilst conceptually simple, image-based registration via homographies might not be applicable to the changing surgical scene. However, for the purpose of imitation learning, in this work, and as will be discribed in detail in Chapter 5, we are only predicting homographies incrementally. This assumption implies that the surgical scene is temporally unchanged, which holds to good extend. Homographies are thus a good candidate for describing actions $\hat{a}_t^*$, necessary for Fig. 1.19. They additionally can be utiliezd to feedback a control signal to the robot. In this section, we will first describe homographies and their theoretical background. Next, we will explain how they can be used to control the velocity of a camera frame C.

Homographies and Four Point Representation Two images are related by a homography if both images view the same plane from different angles and distances. Points on the plane, as observed by the camera from different angles in homogeneous coordinates $\mathbf{p}_i = \begin{bmatrix} u_i & v_i & 1 \end{bmatrix}^{\mathrm{T}}$ are related by a projective homography $\mathbf{G}$ [Malis 2007]

$$\alpha_g \mathbf{p}_i = \mathbf{G}\mathbf{p}_i'. \tag{1.11}$$

Since the points $\mathbf{p}_i$ and $\mathbf{p}_i'$ are only observed in the 2D image, depth information is lost, and the projective homography $\mathbf{G}$ can only be determined up to scale α_g. The distinction between projective homography $\mathbf{G}$ and homography in Euclidean coordinates $\mathbf{H} = \mathbf{K}^{-1}\mathbf{G}\mathbf{K}$, with the camera intrinsics $\mathbf{K}$, is often not made for simplicity, but is nonetheless important for control purposes. The eight unknown parameters of $\mathbf{G}$ can be obtained through a set of $N \geq 4$ matching points $\mathbb{P} = \{(\mathbf{p}_i, \mathbf{p}_i'), i \in [0, N-1]\}$ by rearranging (1.11) into

$$\begin{bmatrix} u_i' & v_i' & 1 & 0 & 0 & 0 & -u_i'u_i & -v_i'u_i & -u_i \\ 0 & 0 & 0 & u_i' & v_i' & 1 & -u_i'v_i & -v_i'v_i & -v_i \end{bmatrix} \mathbf{g} = \mathbf{0} \quad \forall i, \tag{1.12}$$

where $\mathbf{g}$ holds the entries of $\mathbf{G}$ as a column vector. The ninth constraint, by convention, is usually to set $||\mathbf{g}||_2 = 1$. Classically, $\mathbb{P}$ is obtained through feature detectors but it may also be used as a means to parameterise the spatial transformation. Recent deep approaches indeed set $\mathbb{P}$ as the corners of an image, and predict $\Delta\mathbf{p}_i = \mathbf{p}_i' - \mathbf{p}_i$. This is also known as the four point homography $\mathbf{G}_{\text{4point}}$

$$\mathbf{G}_{\text{4point}} = \begin{bmatrix} \Delta u_0 & \Delta v_0 \\ \Delta u_1 & \Delta v_1 \\ \Delta u_2 & \Delta v_2 \\ \Delta u_3 & \Delta v_3 \end{bmatrix}, \tag{1.13}$$

which relates to $\mathbf{G}$ through (1.12), where $\mathbf{p}_i' = \mathbf{p}_i + \Delta\mathbf{p}_i$.

Relation to Camera Frame Velocity Given a projective homography $\mathbf{G}$ in image space, a camera body frame velocity that seeks to minimize $\Delta\mathbf{p}_i$ can be derived à la [Benhimane 2006].

Be $\mathbf{x}' = \begin{bmatrix} x' & y' & z' \end{bmatrix}^T$ a point in 3D coordinates and its observation under a homography transform $\mathbf{H}$:

$$\mathbf{x} = \mathbf{H}\mathbf{x}'. \tag{1.14}$$

In normalized coordinates $\mathbf{m}' = \frac{1}{z'}\mathbf{x}'$ we get

$$\frac{z}{z'}\mathbf{m} = \mathbf{H}\mathbf{m}'. \tag{1.15}$$

It can then be shown that a translational error ${}^{\mathrm{C}}\mathbf{e}_v$ and a rotational error $\left[{}^{\mathrm{C}}\mathbf{e}_\omega\right]_\times$ exist that yield camera body frame velocities which locally converge $\mathbf{m} \to \mathbf{m}'$. These can be expressed through

$$\begin{aligned} {}^{\mathrm{C}}\mathbf{e}_v &= (\mathbf{H} - \mathbf{I})^{\mathrm{C}'}\mathbf{m}' \\ \left[{}^{\mathrm{C}}\mathbf{e}_\omega\right]_\times &= \mathbf{H} - \mathbf{H}^{\mathrm{T}}. \end{aligned} \tag{1.16}$$

1.6.2.4 Robotic Laparoscope Control

Given the relation between image space action and camera frame velocity that was introduced in the above Section 1.6.2.3, only a mapping from camera frame velocity to joint space velocity $\dot{\mathbf{q}}$ is missing for controlling a robotic laparoscope holder. To this end we assume that the transformation from robot end-effector to camera frame ${}^{E}\mathbf{\Theta}_{C}$ is accurately known. This can e.g. be achieved through the methods described in Section 1.4.1 and Section 1.4.2. Given the kinematics of the serial manipulator and the homogeneous transform end-effector to camera ${}^{E}\mathbf{\Theta}_{C}$, a relation between joint velocties $\dot{\mathbf{q}}$ and camera frame velocity ${}^{B}\dot{\mathbf{x}}_C$ can be established through the Jacobian matrix $\mathbf{J}$ as follows

$${}^{B}\dot{\mathbf{x}}_C = \mathbf{J}\dot{\mathbf{q}}. \tag{1.17}$$

The Jacobian therein can be determined analytically through geometrical considerations. Inverting this locally linear equation yields target joint velocities

$$\dot{\mathbf{q}} = \mathbf{J}^{\dagger\,B}\dot{\mathbf{x}}_C. \tag{1.18}$$

This relation may e.g. be used in a proportional-integral-derivative (PID) control scheme. The exact use of (1.18) to control a camera frame body velocity through (1.16) with additional RCM constraint will further be contributed as part of Chap-

ter 3 in Section 3.2.

Given the necessary preliminary considerations, the following Section 1.7 will outline the structure of this thesis. The subsequent chapters will then delve into the technical details of the proposed approach.

1.7 Thesis Structure

Chapter 1 revisited clinical, economical, and technical considerations that lead to the hypothesis of embodiment-invariant IfO for laparoscopic camera motion automation. The remainder of the thesis will present the technical means to achieve this. Each of these technical chapters, Chapter 2 to Chapter 5, is an *in extenso* reproduction of work we published. The overarching goal is the exploration of Fig. 1.19.

Chapter 2 addresses the need for improved spatial awareness in robotic laparoscopy that was identified as key-requirement for automation in Chapter 1.3.3.1. The desire for marker-free eye-in-hand and eye-to-hand calibration from Section 1.4 will be achieved through the introduction of a novel robust point-to-plane ICP algorithm on a Lie algebra. The proposed method uses the segment anything model (SAM) [Kirillov 2023] and is extensible to any serial arm given its meshes or computer-aided design (CAD) model. A integration into the ROS 2 ecosystem is provided and in-vivo studies on a dual arm system are conducted.

Chapter 3 addresses the dominant tool following assumption of Section 1.5.2. A novel control scheme for robotic laparoscope holders that combines Section 1.6.2.3 and Section 1.6.2.4 through a RCM objective with a homography-based camera frame velocity task $\hat{a}_t$ is introduced. The effectiveness of the proposed control scheme is demonstrated on a phantom setup in a semi-autonomous scenario (level three autonomy, Section 1.1). The locality and temporality assumptions of the proposed method therein are discussed in the context of the proposed IL framework of Fig. 1.19.

Chapter 4 addresses the lack of state-action pairs $(\hat{s}_t, \hat{a}_t^*)$ for IL in laparoscopic camera motion automation. Therefore, a novel supervised method for reliably extracting camera motion from videos of laparoscopic interventions despite object and tool motion is introduced, a necessity for the realization of Fig. 1.19. The method exploits the locking mechanism of the da Vinci® robot, which was explained in Fig. 1.20. Camera motion is synthetically added to the SurgVisDom dataset [Zia 2021] through the introduction of a *homography generation algorithm*. The pro-

posed method is evaluated on the Cholec80 dataset [Twinanda 2017], thus testing generalizability from RMIS to MIS.

Chapter 5 finally addresses the prediction of laparoscopic camera motion from images through IL of an expert policy $\pi_{\mathrm{E}} : \hat{s}_t \rightarrow \hat{a}_t^*$. Therefore, a novel self-supervised training scheme for learning to predict camera motion in the form of homographies from images is introduced. The method builds on the laparoscopic camera motion extraction of Chapter 4 for runtime estimation of camera motion on videos of laparoscopic inverventions. A novel sampling scheme for camera motion importance sampling on vast databases of laparscopic interventions is contributed. Early signs of camera motion predictions capabilites are found on the Cholec80 [Twinanda 2017] and the HeiChole [Maier-Hein 2021] datasets through comparison against baselines. Prediction-label correspondence is demonstrated on the AutoLaparo [Wang 2022] dataset. This work, in combination with Chapter 3, thus indicates that camera motion may be learned on camera assistant-held laparoscopes and transferred to robotic laparoscope holders, as was targeted in Fig. 1.19.

Appendix A provides a description of the software stack and ROS 2 integration for the KUKA LBR Med7/14 and IIWA7/14 series that was developed as part of this thesis. The stack is open-source and freely accesible on GitHub. It plays a key role in the work presented in Chapter 2 and Chapter 3.

Table 1.3: Exhaustive overview of publicly available MIS and RMIS datasets. All datasets were acquired and analyzed for task-appropriate metrics. Datasets that were not available, or are unreasonable for evaluation, are marked with N/A, where datasets were not available or unreasonable to analyze.

Collection	Specifications						
	Name	Year	Length / #	Frame Rate / Hz	Resolution / pixels	Camera Motion	Note
Endoscopic Vision Challenge	MISAW [Mitsuishi 2013]	2020	128102	30	460×540	no	synthetic
	SurgVisDom [Zia 2021]	2020	185620	20	540×960	occasional	da Vinci®
	ROBUST-MIS [Maier-Hein 2021][Ross 2020]	2019	7546968	25	540×960	yes	laparoscopic
	SCARED	2019	16818	25	1024×1280	yes	exoscopic
	SWASA	2019	N/A	N/A	N/A	yes	laparoscopic
	SWASA	2018	N/A	N/A	N/A	yes	laparoscopic
	RSS [Allan 2020]	2018	2235	2	1024×1280	occasional	da Vinci®
	RIS [Allan 2019]	2017	3225	2	1024×1280	occasional	da Vinci®
	KBD	2017	3000	2	1024×1280	occasional	da Vinci®
	ISAT	2015	16243	25	$480 \times 640/576 \times 720$	yes	laparoscopic
Hamlyn Center Datasets	Siamese [Ye 2017]	2017	34240	30	192×384	occasional	da Vinci®
	Giannarou left / right [Giannarou 2013]	2012	8063	30	480×640	occasional	da Vinci®
	Mountney left / right [Mountney 2010]	2010	14418	30	480×640	occasional	da Vinci®
Other	YouTube	N/A	N/A	N/A	N/A	N/A	N/A
	SARAS-ESAD [Bawa 2020]	2020	18793	1	1080×1920	occasional	da Vinci®
	Cholec80 [Twinanda 2017]	2017	4612530	25	480×854	yes	laparoscopic
	JIGSAW [Ahmidi 2017]	2016	527491	30	480×640	no	synthetic

Chapter 2

Marker-free Unified Eye-Hand Calibration

Table of Contents

Disclaimer This Chapter 2 is an *in extenso* reproduction of work prepared for consideration in a journal publication. Only Section 2.1 is altered to highlight additional context within the scope of this thesis and remove redundancy with Chapter 1.

2.1 Introduction

Multi-robot arm systems possess an incredible potential to revolutionize the way we approach complex tasks, improving efficiency and productivity. By utilizing multiple arms, we can overcome the limitations of a single unit, extending the range of possible tasks that the system can perform. For example, suturing requires at least two arms. There exists a wide variety of multi-robot arm applications beyond surgical robotics including agriculture [Xiong 2020], civil engineering [Yasutomi 2023], space robotics [Yan 2020], packaging and assembly [Do 2012], nuclear decommissioning [Bakari 2007], disaster recovery [Kamezaki 2016], and sub-sea robotics [Brantner 2021].

Direct teleoperation, classified as level zero in the taxonomy proposed by Yang et al. [Yang 2017], is a control approach that maps signals from a surgeon-device interface onto robot motion [Niemeyer 2008], typically using position/velocity based commands. Currently, direct teleoperation is the gold-standard for the majority of surgical robotic systems in use today. This approach guarantees that the operating surgeon is solely responsible for the robot's motion, emphasizing the need for skilled training and practice [Liu 2015]. Lack of training, as explained in the introductory Section 1.3.3.1, may lead to severe surgical workflow interruptions. Collision models and collision avoidant control is thus desirable. A prerequisite to collision avoidant control, however, is precise localization. With the increasing automation of surgical procedures, e.g. see recent work by Saeidi et al [Saeidi 2022], the demand for localization strategies and collision models becomes even more pronounced.

A key requirement for future automated systems is robot base-to-base localization, i.e. each robot arm must be aware of each other arm's relative position and orientation. This is required to implement an effective collision model into the automated control system ensuring patient and the operating team's safety. In the case where multi-robot arm systems are deployed on a surgical table, base localization can be derived from CAD models, however, these measurements may not accurately reflect reality due to manufacturing inaccuracies. In the case of the increasingly popular modular systems, mentioned in Section 1.3.2, the arm setup may be different every time the theater is prepared. Furthermore, robotic systems must also undergo localization in relation to other elements present in the operating theater, including patient anatomy, lighting fixtures, imaging systems (e.g. C-arm), and so on, see Fig. 1.4. This motivates the need for a robot base calibration step introduced into the surgical workflow. Such a step, however, should be designed in a way that can be quickly and safely performed by non-robotics experts, i.e. the theater staff.

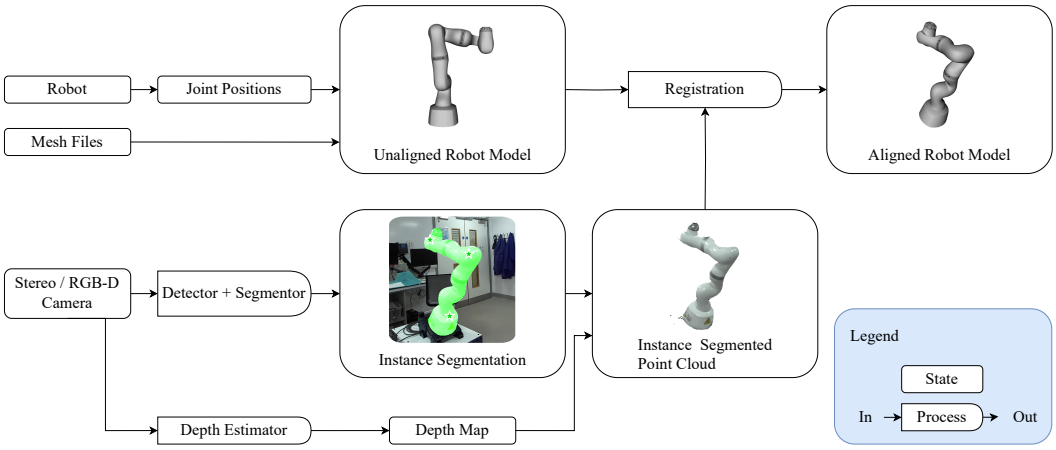


Figure 2.1: Proposed registration procedure. The pipeline takes joint positions and corresponding stereo or RGB-D images as input and yields eye-to-hand transformations. Upper half: Joint positions from the robot(s) are used to transform the link mesh files into an unaligned robot model. Lower half: Stereo (or RGB-D) images are fed through a depth estimator to obtain a depth map. A monocular image is taken from the stereo image and is used to detect and instance segment the robot. The depth map and instance segmentation are fused to obtain an instance-segmented point cloud. The instance-segmented point cloud is registered to the unaligned robot model to obtain a robot model that is aligned with the observed image. The transform from unaligned to aligned robot model describes the robot to camera homogeneous transform ${}^{C}\boldsymbol{\Theta}_{B}$. Refers to Section 2.4.2.

2.1.1 Contributions

In this work we make the following contributions:

- A unified approach for eye-in-hand and eye-to-hand calibration of multiple surgical robots with no need for any calibration targets since the robot itself is used instead. Our proposed method uses a novel robust ICP formulation of a point-to-plane objective on a Lie algebra. Only RGB-D images, the robots' joint positions, and their respective mesh files or CAD models are used.
- We conduct several benchmarks comparing our methods against classical approaches. Our method ensures safety and, based on the comparisons, demonstrates faster execution and better outcomes when compared to other options.
- Hardware realization of the approach on a dual robot arm system. Furthermore, we showcase a human-robot collaboration through admittance control with arm-arm collision avoidance. The collision model for the two robots relies on the measurements generated by our proposed method.
- We conduct an in-vivo experiment in a clinically relevant dual-arm setup for

spine surgery.

2.2 Related Work

2.2.1 Eye-in-hand Calibration

Prior work on the eye-in-hand problem [Horaud 1995; Strobl 2006] require the user to fabricate a marker (e.g. checkerboard or AprilTag [Olson 2011]) with known geometry making it easy to extract a pose from images. In our work, we do not require the user to create such a marker. Instead, we make use of the mesh files or CAD models provided by the manufacturer.

2.2.2 Marker-free Registration

Marker-free eye-to-hand calibration is an active research field with new developments being lead by differentiable rendering frameworks, such as PyTorch3D [Ravi 2020] and Kaolin [Fuji Tsang 2022]. Consequentially, some works utilize these frameworks to optimize for camera poses such that mesh renders match image segmentations [Chen 2023]. Further work proposes a fully self-supervised procedure that learns both, to segment and to estimate the eye-to-hand transformation [Lu 2023]. These works require tailoring to dedicated robot models and depend on fully-blown differentiable rendering pipelines, which makes them difficult to use in practise. Labbé et al. [2021] use classical non-differentiable rendering, where the authors also predict joint states on top of the eye-to-hand registration. This work also requires training to dedicated systems. Simpler approaches suggest to predict robot joint positions and extract eye-to-hand registrations through solving a PnP problem [Lee 2020]. In this work, we suggest a conceptually much simpler approach that works for any robot, but uses RGB-D images, and, as some of the other methods, robot joint states. Closest to our work comes [Horváth 2017], where the authors also perform point cloud registration. Their method, however, only registers a single configuration and performs complex clustering of the observed point cloud. The only marker-free work on a surgical system is presented in [Pachtrachai 2016], where the authors utilize the surgical tools as calibration target. Whilst being quite appealing, this method does not allows for external collision avoidance nor hybrid setups.

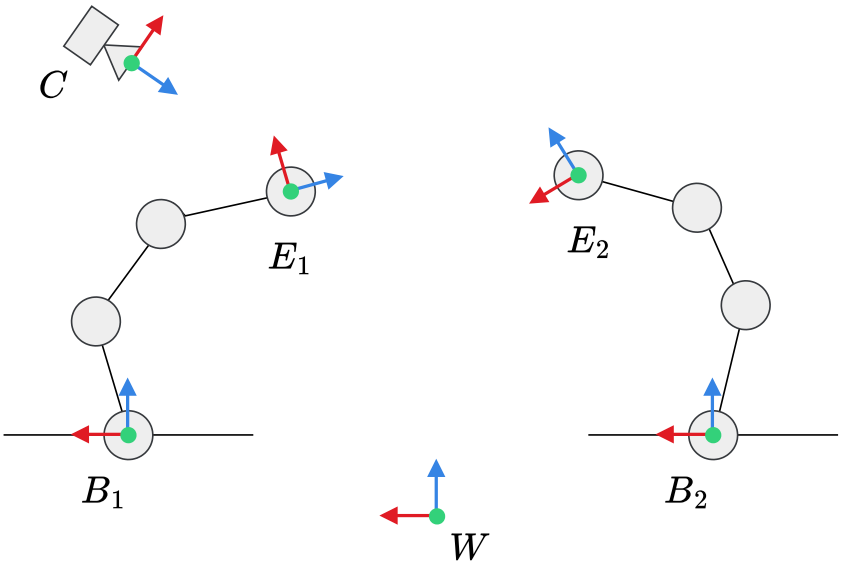


Figure 2.2: Schematic overview of the key coordinate frames of interest for this work. Base-to-base calibration is achieved via two eye-to-hand calibrations. Whilst we show a dual arm system, it is easy to extend the methods discussed in this paper to multiple robots.

2.3 Problem Formulation

In this section, we formally present the two problems that our research addresses. The key coordinate frames are illustrated in the schematic diagrams shown in Fig. 1.11a, and Fig. 1.11b, for the eye-to- and eye-in-hand scenarios, respectively. Fig. 2.2 adds additional nomenclature for the base-to-base via dual eye-to-hand calibration. The following subsection introduces the notation employed in the entirety of this paper and also states some important assumptions we make to facilitate our approach. Subsequently, we outline our problems in the remaining subsections.

2.3.1 Notation and Assumptions

In general, we use letters (e.g. a, α) to denote scalars, bold letters (e.g. $\mathbf{a}, \boldsymbol{\alpha}$) to denote vectors, capital bold letters (e.g. $\mathbf{A}, \boldsymbol{\Gamma}$) to denote matrices, and calligraphy and blackboard formats (e.g. $\mathcal{A}, \mathbb{B}$) to denote sets and spaces.

Coordinate Frame Transformations A homogeneous transformation matrix, denoted ${}^{B}\boldsymbol{\Theta}_{A} \in \mathbb{R}^{4\times4}$, represents the pose of frame A with respect to B (i.e. B, in

this case, is the base frame). The matrix ${}^{B}\mathbf{\Theta}_{A}$ can be expressed as

$$ {}^{B}\mathbf{\Theta}_{A} = \begin{bmatrix} \mathbf{R}({}^{B}\mathbf{q}_{A}) & {}^{B}\mathbf{p}_{A} \\ \mathbf{0}^{1\times 3} & 1 \end{bmatrix} \tag{2.1} $$

where ${}^{B}\mathbf{p}_{A} \in \mathbb{R}^3$ is a column vector representing the position, $\mathbf{R} : \mathbb{R}^4 \to \mathbb{R}^{3\times 3}$ is function that converts the unit-quaternion ${}^{B}\mathbf{q}_{A} \in \mathbb{R}^4$ to a rotation matrix, and 0_3 is the three element row vector containing only zeros. With reference to the schematic of a robot arm in Fig. 1.11b, the camera frame C is expressed in the end-effector frame E by ${}^{E}\mathbf{\Theta}_{C}$, and E is expressed in the robot base frame R by ${}^{R}\mathbf{\Theta}_{E}$.

Forward Kinematics For an N-DoF robot manipulator, the joint positions are denoted $\mathbf{x} \in \mathbb{R}^N$. In this work, we assume that the end-effector frame E expressed in the robot base frame R, i.e. ${}^{R}\mathbf{\Theta}_{E}$, can be computed accurately using forward kinematics. We denote the relationship between the joint positions and link poses, including the end-effector, by

$$ \begin{pmatrix} {}^{R}\mathbf{p}_{L} \\ {}^{R}\mathbf{q}_{L} \end{pmatrix} = \phi(\mathbf{x}), \tag{2.2} $$

where $\phi : \mathbb{R}^N \to \mathbb{R}^7$ represents the forward kinematics function. Given a geometric description of the robot's joints and links (typically in the common URDF format), the forward kinematics $\phi(\cdot)$ is easily derived.

Modality Prerequisites Regarding the camera, that is either placed in the environment, as in Fig. 2.2, or mounted at the robot end-effector, as in Fig. 1.11b, we assume stereo or RGB-D. Finally, we assume access to the mesh files or CAD models for each of the robot arms.

2.4 Materials and Methods

This section introduces a baseline base-to-base calibration procedure via handshake in Section 2.4.1. For the eye-in/to-hand scenarios, baselines as introduced in Section 1.4.2 are utilized. The baseline calibration procedures are considered the reference for the proposed calibration method. The proposed calibration method is explained outlined in Section 2.4.2 and an overview is given in Fig. 2.1 and Fig. 2.3. Instantiations thereof, through a simple ICP, and our novel robust point-to-plane ICP, are detailed in Section 2.4.3 and Section 2.4.4, respectively.

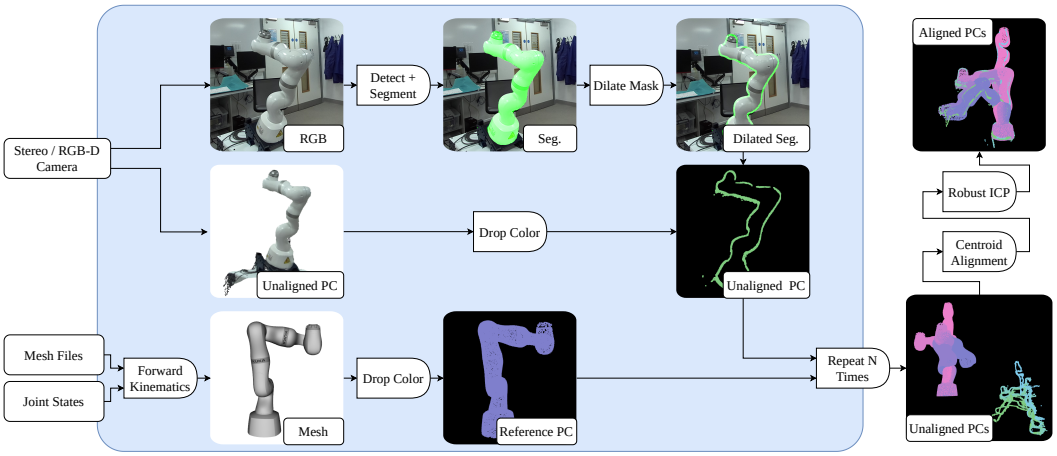


Figure 2.3: Detailed point cloud generation and acquisition including pre-processing steps for the proposed robust point-to-plane ICP. Refers to Section 2.4.4.

In the following sections, homogeneous transforms, from the coordinate frame A to B, are denoted through Equation 2.1 The world coordinate frame is indicated by W. The world coordinate frame W is considered to coincide with the robot base coordinate frame B. W and B may thus be used interchangeably. The camera coordinate frame is highlighted through C. Furthermore, A indicates the coordinate frame of an ArUco marker.

2.4.1 Base-to-base Calibration Baseline

For the handshake calibration procedure, the two robots are rigidly linked to one another at their end-effectors via a calibration fixture of estimated transform ${}^{E_1}\boldsymbol{\Theta}_{E_2}$, see Fig. 2.4a. The robots are then moved in admittance control mode and L base-to-end-effector samples are collected for both robots, i.e. ${}^{B_1}\boldsymbol{\Theta}^{l}_{E_1}$ and ${}^{E_2}\boldsymbol{\Theta}^{l}_{B_2}$. Since the system forms a closed kinematic chain, the base-to-base transform ${}^{B_2}\boldsymbol{\Theta}_{B_1}$ can then be obtained by minimizing:

$$\min_{{}^{E_1}\boldsymbol{\Theta}_{E_2},\,{}^{B_2}\boldsymbol{\Theta}_{B_1}} \sum_{l=0}^{L-1} {}^{B_1}\boldsymbol{\Theta}^{l}_{E_1}\,{}^{E_1}\boldsymbol{\Theta}_{E_2}\,{}^{E_2}\boldsymbol{\Theta}^{l}_{B_2}\,{}^{B_2}\boldsymbol{\Theta}_{B_1} - \mathbf{I}^{4\times4}, \tag{2.3}$$

where $\mathbf{I}^{4\times4}$ is the 4×4 identity matrix. Therein, ${}^{E_1}\boldsymbol{\Theta}_{E_2}$ is part of the optimization to correct for manufacturing errors.

2.4.2 Proposed Registration Procedure

A high-level overview of the proposed method is shown in Fig. 2.1. The method provides an efficient way to find the homogeneous transformation from the robot

base frame to the camera frame ${}^{C}\mathbf{\Theta}_{B}$ and is so appealing over e.g. PnP, since it is directly applicable to any robot given its CAD model. It aligns an unaligned robot model to an observed instance-segmented point cloud through registration. Importantly, the instance-segmented point cloud and unaligned robot model must be synchronized. The unaligned robot model and the instance segmentation are detailed in Section 2.4.2.1.

For the registration step in Fig. 2.1, we investigate a simple ICP algorithm and derive a robust point-to-plane ICP on a Lie algebra. The simple ICP variant of the registration procedure is explained in Section 2.4.3 and the robust point-to-plane ICP variant in Section 2.4.4. A single snapshot is sufficient to find the transformation, which differentiates it from classical calibration procedures described in Section 2.4.1. In the case of a robot-mounted camera, i.e. eye-in-hand calibration illustrated in Fig. 2.4b, the calibration can be used to determine the transformation from robot end-effector to camera ${}^{C}\mathbf{\Theta}_{E}$. In the case of an externally mounted camera, i.e. eye-to-hand calibration illustrated in Fig. 2.4c, the calibration can be used to determine the robot base-to-base transformation ${}^{B_2}\mathbf{\Theta}_{B_1}$ through ${}^{B_2}\mathbf{\Theta}_{B_1} = {}^{C}\mathbf{\Theta}_{B_2}^{-1}\,{}^{C}\mathbf{\Theta}_{B_1}$.

2.4.2.1 Unaligned Robot Model and Instance-segmented Point Cloud

Unaligned Robot Model To obtain an unaligned robot model, the proposed registration procedure takes as input measured joint positions $\mathbf{x}_t$ and link mesh files $\mathcal{M} = \{M_i\}$. Only the mesh vertices $\mathcal{V} = \{\mathbf{V}_i\}$ are extracted from the meshes and everything else, e.g. color, is discarded. All mesh vertices V_i are then transformed through ${}^{L_i}\mathbf{\Theta}_{R}$, where, given the joint positions, ${}^{L_i}\mathbf{\Theta}_{R}$ is obtained via forward kinematics $\phi(\mathbf{x}_t)$, see Equation 2.2. From the transformed mesh vertices $\mathcal{V}$, we sample a total of N points. Since small meshes in CAD files often contain disproportionately many vertices, we normalize the sampled vertices per link by volume. Therefore, we estimate the volume that each link occupies through its respective bounding box volume $b(\mathbf{V}) : \mathbb{R}^M \to \mathbb{R}$. We then sample n_i points per link vertices $\mathbf{V}_i$ via a Poisson-disk sampling strategy, where

$$n_i = N \frac{b(\mathbf{V}_i)}{\sum_j b(\mathbf{V}_j)}. \tag{2.4}$$

This sampling strategy guarantees somewhat equally distributed points in space, although future version might only want to perform Poisson-disk sampling over the entire robot.

Instance-segmented Point Cloud The instance-segmented point cloud is generated by fusing an instance segmentation $\mathbf{S}_t$ with a depth map $\mathbf{D}_t$ as shown in Fig. 2.1. For the instance segmentation, we combine a detector with a segmentor. In this work, a human-in-the-loop acts as the detector by selecting a few points $\mathcal{F}_t$ on the robot in the image. We then deploy the pre-trained SAM [Kirillov 2023] and prompt it with the detection points $\mathcal{F}_t$. We thus achieve an instance segmentation $\mathbf{S}_t$ of the image $\mathbf{I}_t$ with zero-shot training. The depth map $\mathbf{D}_t$ can either be obtained via a stereo camera or an RGB-D camera by processing it with a depth estimator. In this work, we utilize Stereolab's Neural Depth perception as the depth estimator. Since both, the image segmentation $\mathbf{S}_t$ and the depth map $\mathbf{D}_t$, are observed from the same camera coordinate system, one may simply keep depth values where there is a segmentation and discard all others, i.e. compute the Hadamard product. This procedure then provides the instance segmented point cloud.

2.4.3 Simple ICP Registration

In the case of the simple ICP we perform single instance registration. The goal is to align the unaligned robot model (as detailled in Section 2.4.2.1), with the instance-segmented point cloud (as detailled in Section 2.4.2.1), through finding the transform ${}^{C}\mathbf{\Theta}_{B}$. First, the unaligned robot model $\mathcal{V}_\tau$ is generated (as detailled in Section 2.4.2.1). Next, robot instances are detected and segmented, which results in the instance-segmented point cloud $\mathbf{P}_\tau$. We then compute the instance-segmented point cloud's center of mass for an initial guess of the transform ${}^{C}\mathbf{\Theta}_{B,\text{init}}$. Finally, we run an ICP [Glira 2015] algorithm between vertices $\mathcal{V}_\tau$ and point cloud $\mathbf{P}_\tau$, see Equation 2.5. In contrast to [Glira 2015], we only consider vertices from $\mathcal{V}_\tau$ whose surface normals point towards the camera. That is, they should be visible to the camera.

$$\text{ICP} = \underset{{}^{C}\mathbf{\Theta}_{B}}{\arg\min} \, || {}^{C}\mathbf{\Theta}_{B}(\mathcal{V}_\tau) - \mathbf{P}_\tau ||_2 \tag{2.5}$$

2.4.4 Robust Point-to-plane ICP Registration: A Lie Algebra Formulation

To guarantee robust convergence of the noisy point cloud to the robot mesh, we further propose a robust point-to-plane ICP formulation. We capture multiple configurations of the robot. An overview of the method is provided in Fig. 2.3. Deviating from the simple ICP, Section 2.4.3, we dilate the instance segmentations $\mathbf{S}_t$ and only keep the boundary regions. Consequentially, we also only keep the

Algorithm 1 Overview of the robust Lie group formulation of the point-to-plane ICP. Refers to Section 2.4.4.

Let $\{\mathbf{p}\}$ be points to match (the observations). We omit the indices for simplicity here
Let $\{\mathbf{q}, \mathbf{n}\}$ be points with associated normals (the model)
Let $t \leftarrow 0$ be a counter for the outer iterations
Let $\mathbf{\Theta}_0 = \{\mathbf{R}_0, \mathbf{t}_0\}$ be an initial estimate of the transformation parameters
while ICP not converged **do**
    Compute the transformed points $\mathbf{\Theta}_t \mathbf{p}$
    Perform an approximate nearest neighbour search to match $\mathbf{\Theta}_t, \mathbf{p}$, and $\mathbf{q}$ points
    Discard point pairs associated with distances that are definitely outliers
    We now use i indices to indicated matching points
    Compute the residuals: $b_i \leftarrow \tilde{\mathbf{n}}_i^T(\tilde{\mathbf{q}}_i - \mathbf{R}_t\tilde{\mathbf{p}}_i - \mathbf{t}_t)$
        ▷ Previous computation of $\mathbf{\Theta}_t \mathbf{p}$ should of course be re-used here
    Compute the robust standard deviation: $\sigma \leftarrow \text{MAD}(\{b_i\})/0.6745$
        ▷ Done once to avoid recomputing it in the inner iterations
    Let $\tau \leftarrow 0$ be a counter for the inner iterations
    Let $\mathbf{\Theta}_{t,0} \leftarrow \mathbf{\Theta}_t$ be the current estimate of the transformation
    while Gauss-Newton not converged **do**
        if $\tau \neq 0$ **then**
            Compute the residuals: $b_i \leftarrow \tilde{\mathbf{n}}_i^T(\tilde{\mathbf{q}}_i - \mathbf{R}_{t,\tau}\tilde{\mathbf{p}}_i - \mathbf{t}_{t,\tau})$
        end if
        Compute the robust weights: $\omega_i \leftarrow \omega_{1.345\sigma}(b_i)$
        Compute the linear coefficients: $\mathbf{a}_i \leftarrow [-\tilde{\mathbf{n}}_i^T\mathbf{R}_{t,\tau}\tilde{\mathbf{p}}_{i\times}, \tilde{\mathbf{n}}_i^T\mathbf{R}_{t,\tau}]$
        Solve the weighted normal equation:
        $\Delta\mathbf{\Theta} \leftarrow \texttt{chol\_solve}(\mathbf{A}^T\mathbf{W}\mathbf{A}, \mathbf{A}^T\mathbf{W}\mathbf{A})$
        Update the transformation: $\mathbf{\Theta}_{t,\tau+1} \leftarrow \mathbf{\Theta}_{t,\tau} \cdot \exp(\Delta\mathbf{\Theta})$
        $\tau \leftarrow \tau + 1$
    end while
    Let $\mathbf{\Theta}_{t+1} \leftarrow \mathbf{\Theta}_{t,\tau}$
    $t \leftarrow t + 1$
end while

boundary regions of the instance segmented point clouds $\mathcal{P} = \{\mathbf{P}_t\}$. Similarly to the simple ICP, we compute the centroids for the instance segmented point clouds $\{\mathbf{P}_t\}$ as well as for each robot model $\{\mathcal{V}_t\}$. We then use the Kabsch algorithm to find an initial estimate for the camera to base homogeneous transformation through centroid alignment. Given the initial alignment, we then perform a robust point-to-plane registration on a Lie algebra, as further detailed below.

Basic Formulation We first introduce the problem solved by the point-to-plane ICP algorithm [Rusinkiewicz 2001]. Let $\mathbf{p}_i$ and $\mathbf{q}_i$ be corresponding 3D points in homogeneous coordinates. Let $\mathbf{n}_i$ be a normal vector associated with $\mathbf{q}_i$. We search a rigid body transformation $\mathbf{\Theta}$ in $SE(3)$ that minimizes:

$$\sum_i ||\mathbf{n}_i^T \cdot (\mathbf{\Theta} \cdot \mathbf{p}_i - \mathbf{q}_i)||^2 = \sum_i (\mathbf{n}_i^T \cdot \mathbf{\Theta} \cdot \mathbf{p}_i - \mathbf{n}_i^T \cdot \mathbf{q}_i)^2 \tag{2.6}$$

Iterative Optimization on the $SE(3)$ Lie Group Let $\mathbf{\Theta}_0$ be the current estimate. In a Lie group iterative optimization approach [Mahony 2002; Vercauteren 2007], we seek a perturbation $\Delta\mathbf{\Theta} \in \mathfrak{se}(3)$ composed with $\mathbf{\Theta}_0$: $\mathbf{\Theta}_1 = \mathbf{\Theta}_0 \cdot \exp(\Delta\mathbf{\Theta})$. We now seek to minimize:

$$\sum_i (\mathbf{n}_i^T \cdot \mathbf{\Theta}_0 \cdot \exp(\Delta\mathbf{\Theta}) \cdot \mathbf{p}_i - \mathbf{n}_i^T \cdot \mathbf{q}_i)^2 \tag{2.7}$$

Mathematical Preliminaries For convenience, we take advantage of the relationship between the 3D cross product, the Lie bracket on SO(3), and the skew-symmetric matrix operator. Let $\boldsymbol{\omega} = [\omega_x, \omega_y, \omega_z]^T$ and $\boldsymbol{\eta} = [\eta_x, \eta_y, \eta_z]^T$ be two 3D vectors, we have

$$[\boldsymbol{\omega}, \boldsymbol{\eta}] = \boldsymbol{\omega} \times \boldsymbol{\eta} = \boldsymbol{\omega}_\times \boldsymbol{\eta} \tag{2.8}$$

where

$$\boldsymbol{\omega}_\times = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix} \tag{2.9}$$

It can be shown that the Lie group exponential in SO(3) admits a closed form through the Rodrigues' formula. Let us consider $\boldsymbol{\omega}$ as an element of $\mathfrak{so}(3)$, we now have

$$\exp_{\text{Lie}}(\boldsymbol{\omega}) = \exp(\boldsymbol{\omega}_\times) = \mathbf{I}^{3\times 3} + \frac{\sin(||\boldsymbol{\omega}||)}{||\boldsymbol{\omega}||}\boldsymbol{\omega}_\times + \frac{1 - \cos(||\boldsymbol{\omega}||)}{||\boldsymbol{\omega}||^2}\boldsymbol{\omega}_\times^2 \tag{2.10}$$

which for small values of $||\boldsymbol{\omega}||$ leads to the following numerically stable approximation

$$\exp_{\text{Lie}}(\boldsymbol{\omega}) \approx \mathbf{I}^{3\times 3} + (1 - \frac{||\boldsymbol{\omega}||^2}{6} + \frac{||\boldsymbol{\omega}||^4}{120})\boldsymbol{\omega}_\times + (\frac{1}{2} - \frac{||\boldsymbol{\omega}||^2}{24} + \frac{||\boldsymbol{\omega}||^4}{720}))\boldsymbol{\omega}_\times^2 \tag{2.11}$$

Let us consider a 6D vector $\mathbf{u} = [\boldsymbol{\omega}; \boldsymbol{\tau}]$ as an element of $\mathfrak{se}(3)$. Its matrix representation is provided by

$$\mathbf{u}_\dagger = \begin{bmatrix} \boldsymbol{\omega}_\times & \boldsymbol{\tau} \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{4\times 4} \tag{2.12}$$

which leads to the following closed-form formula for the Lie group exponential for SE(3):

$$\exp_{\text{Lie}}(\mathbf{u}) = \exp(\mathbf{u}_\dagger) = \begin{bmatrix} \exp_{\text{Lie}}(\boldsymbol{\omega}) & L(\boldsymbol{\omega})\boldsymbol{\tau} \\ 0 & 1 \end{bmatrix} \tag{2.13}$$

where

$$L(\boldsymbol{\omega}) = \mathbf{I}^{3\times 3} + \frac{1 - \cos(||\boldsymbol{\omega}||)}{||\boldsymbol{\omega}||^2}\boldsymbol{\omega}_\times + \frac{||\boldsymbol{\omega}|| - \sin(||\boldsymbol{\omega}||)}{||\boldsymbol{\omega}||^3}\boldsymbol{\omega}_\times^2 \tag{2.14}$$

where a Taylor expansion should again be used for numerical stability if $||\boldsymbol{\omega}||$ is small:

$$L(\boldsymbol{\omega}) \approx \mathbf{I}^{3\times 3} + (\frac{1}{2} - \frac{||\boldsymbol{\omega}||^2}{24} + \frac{||\boldsymbol{\omega}||^4}{720})\boldsymbol{\omega}_\times + (\frac{1}{6} - \frac{||\boldsymbol{\omega}||^2}{120} + \frac{||\boldsymbol{\omega}||^4}{5040})\boldsymbol{\omega}_\times^2 \tag{2.15}$$

First-order Linearization of the Action of the Exponential in $\mathfrak{se}(3)$ Let $\mathbf{p} = [\tilde{\mathbf{p}}; 1]$ be a 3D point in homogeneous coordinates. We consider an infinitesimal element $\Delta\mathbf{u} = [\Delta\boldsymbol{\omega}; \Delta\boldsymbol{\tau}]$ of $\mathfrak{se}(3)$ and consider the corresponding action on p:

$$\begin{aligned} \exp_{\text{Lie}}(\Delta\mathbf{u})\mathbf{p} \approx (\mathbf{I}^{4\times 4} + \Delta\mathbf{u}_\dagger)\mathbf{p} &= \mathbf{p} + \begin{bmatrix} \Delta\boldsymbol{\omega}_\times & \Delta\boldsymbol{\tau} \\ 0 & 0 \end{bmatrix}\mathbf{p} = \mathbf{p} + \begin{bmatrix} \Delta\boldsymbol{\omega}_\times\tilde{\mathbf{p}} + \Delta\boldsymbol{\tau} \\ 0 \end{bmatrix} \\ &= \mathbf{p} + \begin{bmatrix} \Delta\boldsymbol{\omega} \times \tilde{\mathbf{p}} + \Delta\boldsymbol{\tau} \\ 0 \end{bmatrix} = \mathbf{p} + \begin{bmatrix} -\tilde{\mathbf{p}} \times \Delta\boldsymbol{\omega} + \Delta\boldsymbol{\tau} \\ 0 \end{bmatrix} \\ &= \mathbf{p} + \begin{bmatrix} -\tilde{\mathbf{p}}_\times & \mathbf{I}^{3\times 3} \\ 0 & 0 \end{bmatrix}\Delta\mathbf{u} = \mathbf{p} + \mathbf{D}(\mathbf{p})\Delta\mathbf{u} \end{aligned} \tag{2.16}$$

where

$$\mathbf{D}(\mathbf{p}) = \begin{bmatrix} -\tilde{\mathbf{p}}_\times & \mathbf{I}^{3\times 3} \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{4\times 6} \tag{2.17}$$

Lie Algebra Linearization of the Point-to-plane Objective Plugging (2.16) in the original point-to-plane ICP cost function we get the following linearization:

$$\begin{aligned}&\sum_i \Big(\mathbf{n}_i^T \boldsymbol{\Theta}_0(\mathbf{p}_i + \mathbf{D}(\mathbf{p}_i)\Delta\boldsymbol{\Theta}) - \mathbf{n}_i^T\mathbf{q}_i\Big)^2 \\ &\quad = \sum_i \Big(\mathbf{n}_i^T \boldsymbol{\Theta}_0 \mathbf{D}(\mathbf{p}_i)\Delta\boldsymbol{\Theta} + \mathbf{n}_i^T\boldsymbol{\Theta}_0\mathbf{p}_i - \mathbf{n}_i^T\mathbf{q}_i\Big)^2\end{aligned} \tag{2.18}$$

Denoting $\mathbf{a}_i = \mathbf{n}_i^T\boldsymbol{\Theta}_0\mathbf{D}(\mathbf{p}_i)$, $\mathbf{A} = [\mathbf{a}_0; \ldots; \mathbf{a}_N]$, $b_i = \mathbf{n}_i^T(\mathbf{q}_i - \boldsymbol{\Theta}_0\mathbf{p}_i)$, and $\mathbf{B} = [b_0; \ldots; b_N]$, we end up with a standard linear least squares problem:

$$||\mathbf{A}\Delta\boldsymbol{\Theta} - \mathbf{B}||^2 \tag{2.19}$$

which can conveniently be expressed using the pseudo-inverse $\mathbf{A}^\dagger$ of $\mathbf{A}$:

$$\Delta\boldsymbol{\Theta} = \mathbf{A}^\dagger\mathbf{B} \tag{2.20}$$

In practice, given the dimensions of $\mathbf{A}$, b and $\Delta\boldsymbol{\Theta}$, (2.19) is probably best solved by using the normal equations

$$(\mathbf{A}^T\mathbf{A})\Delta\boldsymbol{\Theta} = \mathbf{A}^T\mathbf{B} \tag{2.21}$$

and relying on a Cholesky decomposition:

$$\Delta\boldsymbol{\Theta} = \texttt{chol\_solve}(\mathbf{A}^T\mathbf{A}, \mathbf{A}^T\mathbf{B}) \tag{2.22}$$

Note that by expressing $\mathbf{A}$ and $\mathbf{B}$ without homogeneous coordinates, we get:

$$\mathbf{a}_i = [-\tilde{\mathbf{n}}_i^T\mathbf{R}_0\tilde{\mathbf{p}}_{i\times}, \tilde{\mathbf{n}}_i^T\mathbf{R}_0] \in \mathbb{R}^{1\times 6} \tag{2.23}$$

and

$$b_i = \tilde{\mathbf{n}}_i^T(\tilde{\mathbf{q}}_i - \mathbf{R}_0\tilde{\mathbf{p}}_i - T_0) \in \mathbb{R} \tag{2.24}$$

Robust Formulation To reduce the influence of outliers on the solution, a typical approach is to introduce a loss function ρ_κ parameterized by a scale parameter (a.k.a. soft margin or cut-off) κ to scale the residuals. We then seek to minimize:

$$\sum_i \rho_k\Big(\mathbf{n}_i^T \cdot \boldsymbol{\Theta} \cdot \mathbf{p}_i - \mathbf{n}_i^T \cdot \mathbf{q}_i\Big) \tag{2.25}$$

or equivalently if we start from a given estimate in the above Lie group setting:

$$\sum_i \rho_k\Big(\mathbf{a}_i \Delta\boldsymbol{\Theta} - b_i\Big) \tag{2.26}$$

Using the square loss function $\rho_\kappa(z) = z^2$ leads to the original least-squares problem. A classical robust alternatives is the Huber loss

$$\rho_\kappa(z) = \begin{cases} z^2 & \texttt{if } |z| <= \kappa \\ 2|z|\kappa - \kappa^2 & \texttt{otherwise} \end{cases} \tag{2.27}$$

with $\kappa = 1.345\sigma$ being advocated to provide robustness while retaining appropriate properties when the errors are Gaussian.

This minimisation can be achieved by a standard iteratively reweighted least squares (IRLS) approach [Green 1984] or by including a second order terms in a slightly more advanced reweighted Gauss-Newton approach as discussed in [Triggs 2000]. The latter is referred to as the Triggs correction in [1]. While elegant, previous work has shown that the Triggs correction does not significantly improve on the standard IRLS [Zach 2014; Zach 2018]. Here, for simplicity we thus restrict ourselves to standard IRLS. Given a choice of scaling factor κ and an estimate of the parameters $\boldsymbol{\Theta}$, the robust problem is converted to a weighted least-squares:

$$\sum_i \omega_\kappa(b_i)||\mathbf{a}_i \Delta\boldsymbol{\Theta} - b_i||^2 \tag{2.28}$$

with $\omega_\kappa(z) = \rho'_\kappa(z)/z$ being the weight function associated with the Huber loss:

$$\omega_\kappa(z) = \begin{cases} 1 & \texttt{if } |z| <= \kappa \\ \kappa/|z| & \texttt{otherwise} \end{cases} \tag{2.29}$$

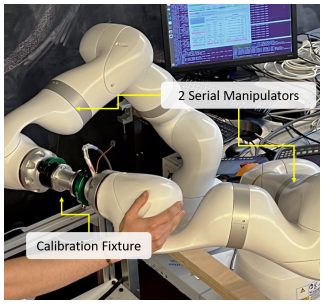
This leads us to the following weighted normal equation:

$$(\mathbf{A}^T\mathbf{W}\mathbf{A})\Delta\boldsymbol{\Theta} = \mathbf{A}^T\mathbf{W}\mathbf{B} \tag{2.30}$$

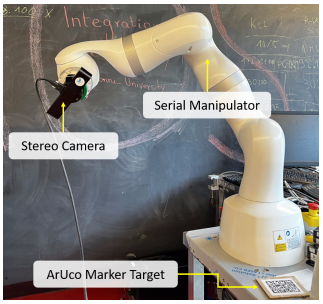
where $\mathbf{W} = \text{diag}\left(\omega_\kappa(b_0), \ldots, \omega_\kappa(b_N)\right)$.

One aspect we haven't addressed yet it the choice of κ for the Huber loss. As mentioned earlier, a typical choice is to use $\kappa = 1.345\sigma$. We are thus left with the need to robustly estimate the standard deviation of the residuals. This is typically

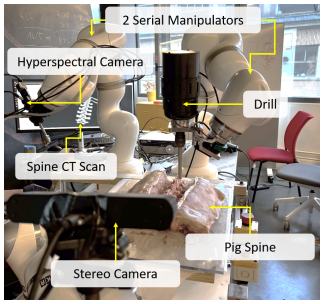
[1] `http://ceres-solver.org/nnls_modeling.html#theory`



(a) Handshake calibration for base-to-base calibration. The two serial manipulators are rigidly attached and in admittance control mode.



(b) Eye-in-hand calibration obtained through self-observation, where the serial manipulator observes itself.



(c) Eye-to-hand calibration in a dual serial manipulator setup to obtain base-to-base calibration.

Figure 2.4: Calibration procedures. The proposed registration procedure, see Fig. 2.1, is evaluated against alternative calibrations. The base-to-base calibration via eye-to-hand registration in (c) is compared against a handshake calibration in (a). The eye-in-hand calibration in (b) is compared against a classical calibration via an ArUco marker target, also (b). Refers to Section 2.5.1.

done through a median absolute deviation (MAD):

$$\hat{\sigma} = \frac{\text{MAD}(\{b_i\})}{0.6745} = \frac{\text{median}(\{|b_i - \text{median}(\{b_i\}|)\})}{0.6745} \tag{2.31}$$

2.5 Experimental Setup

Two types of experiments are considered, ex-vivo experiments in Section 2.5.1, and in-vivo experiments in Section 2.5.2. The ex-vivo experiments include comprehensive comparisons against eye-in-hand, eye-to-hand, and base-to-base baselines for the simple ICP algorithm from Section 2.4.3. An overview for the different ex-vivo calibration baselines is given in Fig. 2.4. This section further includes downstream applications such as the collision avoidant admittance control. Furthermore, an initial ex-vivo dual arm calibration for the proposed robust point-to-plane ICP algorithm from Section 2.4.4 is presented. The in-vivo experiments use the robust point-to-plane ICP algorithm from Section 2.4.4. The in-vivo experimental setup is presented in Fig. 2.5.

2.5.1 Ex-vivo Experiments

The experimental setup is shown in Fig. 2.4a-Fig. 2.4c. We utilize two KUKA LBR Med7 robots and control them via the LBR-Stack (Appendix A) at a control rate of

100 Hz. For the handshake calibration, we deploy them in impedance control mode, for all other calibrations we use position control mode. The robots are mounted on a table and are fixed with respect to each other. By construction, their distance is 38 cm. For the camera we use a ZED 2i (Stereolabs, USA). We collect images and depth maps with a resolution of 448×256 at a frame rate of 15 Hz. For the depth estimation, we utilize the camera in neural depth perception mode.

Four baseline experiments are conducted, two regarding eye-in-hand and two regarding eye-to-hand calibration. For the eye-in-hand calibration, the camera is mounted to the robot via a GRIP G-SHW063 tool connector, see Fig. 2.4b. For the baseline calibration experiments, see Section 2.4.1, we use a 3×4 ArUco marker target of square size 15.77 mm. 1259 image-joint position correspondences are collected. For the proposed method, see Section 2.4.3, we have the robot observe itself through the camera. We collect 1272 image-joint position correspondences, but only use one. For the eye-to-hand calibration, the camera is put on an external tripod, see Fig. 2.4c. 790 image-joint position correspondences are collected, again only one is used. Robot 1 and robot 2 are calibrated and the base-to-base calibration is extracted, see Section 2.4.2. The base-to-base calibration is compared to a handshake calibration, see Fig. 2.4a and Section 2.4.1. For the handshake, the robots are rigidly connected via a calibration fixture. 27700 data points are collected of which 277 are used for the calibration.

Finally, arm-arm collision avoidant admittance control using the simple ICP, and exemplary base-to-base calibrations using the robust point-to-plane ICP are conducted. The collision avoidant admittance control is tested by moving the robots towards each other. The base-to-base calibration is evaluated visually.

2.5.2 In-vivo Experiment

To qualitatively verify the proposed method in a clinically relevant scenario, and to collect data, we conduct an in-vivo study. The primary goal of this experiment, however, is data collection. Given the registration, the recorded data can serve as ground truth for further marker-free registration despite draping. In the study, spine surgery is investigated, and in particular, robot-assisted pedicle screw placement. The pig is put under general anesthesia at all times. An overview of the setup is shown in Fig. 2.5. Two KUKA LBR Med7 R800 are deployed. Notably, both robots are draped, turning the proposed registration procedure unfeasible. Thus, registration is performed prior to draping, and neither the robots, nor the camera are moved thereafter. A ZED 2i stereo camera (Stereolabs, USA) is wall-mounted, so both robots and the surgical scene are in sight. We record stereo images and

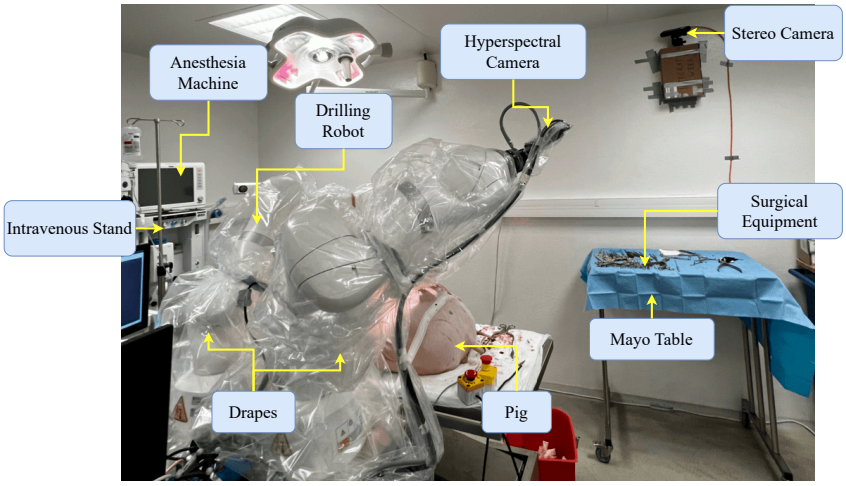


Figure 2.5: Clinical setup. A camera is mounted against a wall and both robots are registered using the proposed method of Section 2.4.4. Notably, the registration is performed prior to draping. Refers to Section 2.5.2.

depth maps at a resolution of 540×960. Both robots are put into admittance control mode and are hand-guided into different configurations for sample collection.

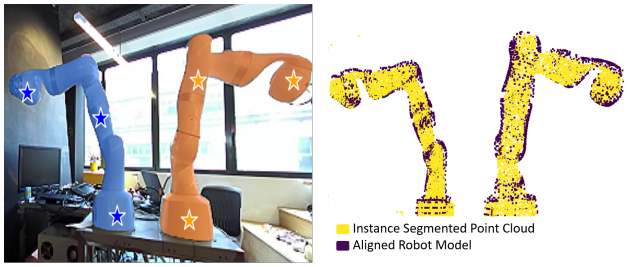
2.6 Results

2.6.1 Ex-vivo Results

Qualitative Baseline Results - Simple ICP The registrations are shown in Fig. 2.6. It can be observed that in both scenarios, i.e. eye-in-hand (Fig. 2.6a) and base-to-base via eye-to-hand (Fig. 2.6b), the segmentation via initial detection works well. The detections $\mathcal{F}_\tau$ are indicated through stars. Robot one and two can be clearly separated. This precise segmentation $\mathbf{S}_\tau$ results in an instance-segmented point cloud with very little outliers. The registration of the unaligned robot models $\mathcal{V}_\tau$ onto the instance-segmented point clouds $\mathbf{P}_\tau$ results in a visually compelling alignment (Fig. 2.6a - right - and Fig. 2.6b - right). Therein, the instance-segmented point clouds are visualized in yellow and the aligned robot models are shown in purple.

Quantitative Baseline Results - Simple ICP Quantitative results are summarized in Table 2.2. For the eye-in-hand calibrations, it can be seen that the classical calibration procedure via ArUco markers deviates significantly from the manufactured values. This is most likely caused by noisy samples. In contrast, the proposed method, although single-shot, compares better with the expected man-

(a) Eye-in-hand calibration, corresponding to Fig. 2.4b. Left: Instance segmentation. The camera is mounted to robot one, highlighted through the blue mask. For visualization purposes, the image is rotated by 180°. Right: Point cloud to robot model registration.

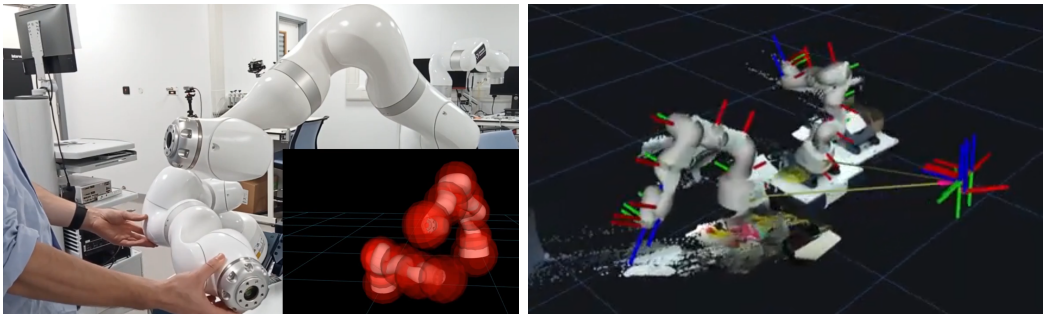


(b) Eye-to-hand calibration, corresponding to Fig. 2.4c. Left: Instance segmentation. The camera is mounted externally and observes robot one (blue mask) and robot two (orange mask). Right: Point cloud to robot model registration.

Figure 2.6: Ex-vivo eye-in-hand and eye-to-hand registrations. The instance-segmented point clouds $\mathbf{P}_\tau$ align well with the robot model $\mathcal{V}_\tau$. Refers to Section 2.6.1.

ufactured values, which aligns well with the qualitative results in Fig. 2.6a. For the base-to-base calibration, it should be noted that the robots are fixed on a plain surface. The manufactured values can, therefore, be considered accurate. The proposed method yields precise results along the y-axis (distance between the robots), whereas the handshake deviates by 1 cm. The proposed method deviates 1 cm along the z-axis from the manufactured and the handshake values. This dimension is the perpendicular offset to the surface plane. It can further be seen that the proposed method deviates slightly from the orientations of the handshake calibrations when compared to the manufactured values. The deviations might be caused by insufficiently synchronized images and joint states.

Qualitative Admittance Control - Simple ICP Given the simple ICP registration results, refer Section 2.4.3, we perform collision avoidant admittance control using the OpTas library [Mower 2023b]. The spherical collision avoidance model is depicted in Fig. 2.7a. Therein, the spherical constraints are indicated through red



(a) Collision avoidant admittance control with spherical constraints, indicated in red.

(b) Base-to-base registration and mesh overlay onto point cloud.

Figure 2.7: Downstream applications of the eye-to-hand calibration. Refers to Section 2.6.1. Videos are made available online[2].

spheres are the links. As seen in the accompanying video (Fig. 2.7a), the robot can be safely hand-guided in admittance control mode whilst avoiding collision within the spherical constraints.

Qualitative Base-to-base Registration - Robust ICP The qualitative results for the base-to-base via dual eye-to-hand calibration using the robust point-to-plane ICP of Section 2.4.4 are shown in Fig. 2.7b. In the accompanying video (Fig. 2.7b), the robot meshes align well with the observed point cloud. Given these results, it is justified to deploy the method in the in-vivo scenario, see Section 2.5.2 and Section 2.6.2.

2.6.2 In-vivo Results

Initially, we collect 12 measurements for the drilling robot, and 14 measurements for the hyperspecral camera robot. These measurements include joint state, image, and depth correspondences. The positioning of both robots with respect to the camera is shown in Fig. 2.5. Using these measurements, we perform a eye-to-hand calibration via the robust point-to-plane ICP, refer Section 2.4.4.

For both robots, we collect a total of 4588 stereo frames and depth maps with corresponding joint states, corresponding to about 10 minutes of data at an average recorded frame rate of 8 fps. We collect an additional 5611 correspondences for only one robot, either the drilling or the hyperspectral camera robot. This corresponds

[2] Collision avoidant admittance control: `https://drive.google.com/file/d/14IhWxNsEmjGmVBnxTaMYIDQDOuo6Ihak/view?usp=sharing`
Base-to-base registration: `https://drive.google.com/file/d/18KFJyxDw6UiQ1N95tSQadi4TyGtBAQvc/view?usp=sharing`

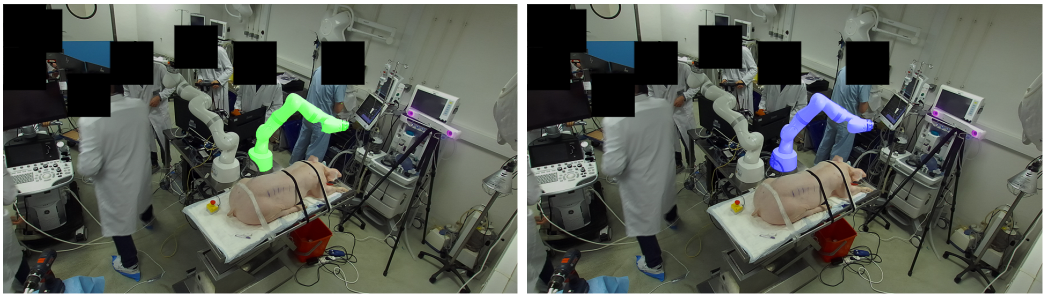
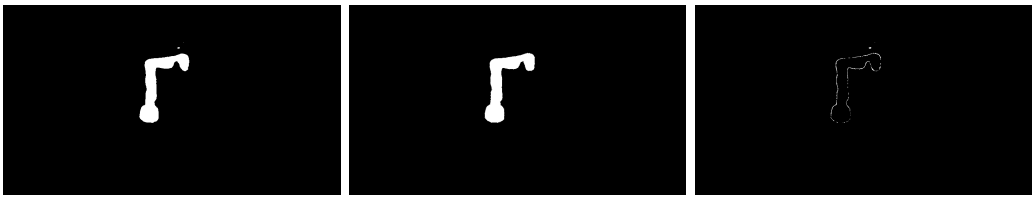


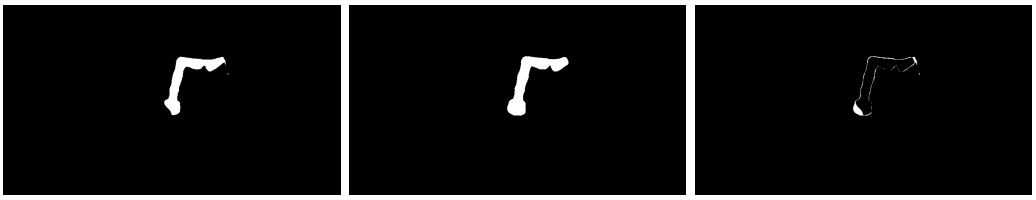
Figure 2.8: Segmented robot (left) and rendered robot given registration (right). Visually, the render shows good alignment with the robot, hinting to an accurate calibration. Refers to Section 2.6.2.



(a) Camera robot: Segmentation.

(b) Camera robot: Rendered mask.

(c) Camera robot: L1-norm of segmentation and render.



(d) Drilling robot: Segmentation

(e) Drilling robot: Rendered mask.

(f) Drilling robot: L1-norm of segmentation and render.

Figure 2.9: Example of segmented and rendered masks given the proposed registration procedure. Refer to Fig. 2.5 for nomenclature. Refers to Section 2.6.2.

to about 7 minutes of data at an average recorded frame rate of 13 fps.

Qualitative results of the registration for the drilling robot are shown in Fig. 2.8. Therein, and given the registration, we render the robot's meshes into the observed image. The render aligns well with the robot. It should be noted that the segmentation, Fig. 2.8 - left, does not cover the entire robot but misses the bottom left bit. This was the case for all observed drilling robot mask and could be corrected through proper selection of points $\mathcal{F}_t$. For the camera robot on the other hand, we obtain close to perfect instance segmentations $\{\mathbf{S}_t\}$. Further exemplary segmentation masks and rendered masks, as well as difference images between segmentation and render, for both, the drilling robot and the camera robot, are shown in Fig. 2.9.

Table 2.1: Average IoU of segmented and rendered mask. Also compare to Fig. 2.9. Refers to Section 2.6.2.

Robot	IoU [a.u]
Camera (left)	0.92 ± 0.01
Drilling (right)	0.84 ± 0.01

For the total set of collected samples, we determine the average IoU between instance segmentations $\mathbf{S}_t$ and renders $\mathbf{R}_t$. Therein, the IoU is expressed through

$$\text{IoU} = \frac{|\mathbf{S}_t \cup \mathbf{R}_t|}{|\mathbf{S}_t \cap \mathbf{R}_t|} \tag{2.32}$$

The results are summarized in Table 2.1. As can be see, and as already visually observed in Fig. 2.9, we find close to perfect average IoU. For the drilling robot, the average IoU is slightly worse than for the camera robot. This is because the instance segmentations $\mathbf{S}_t$ are systematically off by human annotator. Ideally, one should further investigate an error in Cartesian-space, in addition to pixel-space, to determine the true accuracy of the registration. Ground-truth data, however, is not available for this experiment.

2.7 Conclusion and Future Work

In this work we present a novel vision-based robot calibration procedure that unifies eye-in-hand and eye-to-hand calibration through casting them as the same problem formulation. This is achieved by treating the robot as calibration target itself. In the eye-in-hand calibration scenario, the robot simply observes itself whilst in the eye-to-hand scenario the camera observes the robot from an external stand. The introduced formulation can further be used for rapid robot base-to-base calibrations.

Ex-vivo Summary Promising qualitative results for the simple ICP (Section 2.4.3) are presented in Fig. 2.6. For both, eye-in-hand Fig. 2.6a and eye-to-hand Fig. 2.6b, the instance-segmented point cloud, as well as the robot model align well after registration. These qualitative observations also solidify through our quantitative comparisons against baselines, which are summarized in Table 2.2. Whilst an error remains, the proposed method is much quicker, as it is single shot, is much safer to use, as opposed to the handshake calibration, and does not require any calibration targets, as opposed to the eye-in-hand calibration with ArUco target.

In-vivo Summary For the in-vivo experiments Section 2.5.2, we deploy an improved version of the registration algorithm, see Section 2.4.4. Applicability to a clinical scenario was demonstrated, see Fig. 2.5. Visually, we find close to pixel-perfect registrations, see Fig. 2.8. These visual observations are further confirmed through the measured average IoU, see Table 2.1. Whether the amount of collected data is sufficient for marker-free registration despite draping remains to be seen in future work.

Limitations Shortcomings of this work are that a sufficient view of the robot is required to perform the calibration. In a surgical scenario, draping would cause an insufficient view. Consequentially, the calibration would have to be carried out as initialization and the robots would not be allowed to move afterwards. Clinically, however, it is necessary to drape the robots prior to moving them into position, i.e. prior to moving them into the sterile field. This work furthermore only considers robot-robot collision with full knowledge of the robots' states. Robot-staff / surgeon / equipment collisions would require a more general understanding of the scene, potentially even motion prediction.

Future Work In this work we present a very simplistic registration procedure which already provides highly accurate registration results. These results could immediately be deployed to industrial robots, e.g. for collision avoidance purposes. In the clinical scenario, however, draping is an unsolved issue. Since draping does deform the observed point clouds significantly, the presented method would ultimately fail. Methods that would attempt to remove the effects of the draping from the obeserved point cloud are difficult in practise. However, vast amounts of data and precise localizations were collected as part of the in-vivo study, see Section 2.6.2. Future work could, therefore, use this data to develop marker-free registration procedures that might function despite draping in image space, since the draping is somewhat transparent.

Table 2.2: Calibration results using the proposed method (Section 2.4.2) and the baseline methods (Section 2.4.1). In addition to the calibration baselines, the table also lists manufactured values. The transforms are displayed in terms of translations $t_{x/y/z}$ and rotations in terms of Euler angles $r_{x/y/z}$.

Calibration Class	Method	Transform	t_x [m]	t_y [m]	t_z [m]	r_x [°]	r_y [°]	r_z [°]
Eye-in-hand	Manufactured	${}^{C}\mathbf{\Theta}_{E}$	0.0	−0.06	−0.06	0.0	−2.9	−145.0
	ArUco	${}^{C}\mathbf{\Theta}_{E}$	0.0	−0.04	0.03	−1.1	−15.0	−167.7
	Proposed	${}^{C}\mathbf{\Theta}_{E}$	0.0	−0.07	−0.08	1.6	1.6	−146.7
Base-to-base	Manufactured	${}^{B_2}\mathbf{\Theta}_{B_1}$	0.0	−0.38	0.0	0.0	0.0	0.0
	Handshake	${}^{B_2}\mathbf{\Theta}_{B_1}$	−0.01	−0.37	0.0	1.3	0.3	−0.6
Base-to-base via eye-to-hand	Proposed	${}^{B_2}\mathbf{\Theta}_{B_1} = {}^{C}\mathbf{\Theta}_{B_2}^{-1}\,{}^{C}\mathbf{\Theta}_{B_1}$	−0.01	−0.38	−0.01	1.7	2.2	−5.1

Chapter 3

Semi-autonomous Robotic Laparoscope under Remote Center of Motion Constraint

Table of Contents

Disclaimer This Chapter 3 is an *in extenso* reproduction of [Huber 2021]. Only Section 3.1 was altered to highlight additional context within the scope of this thesis.

3.1 Introduction

As was discussed in Section 1.5.1, vision-based automation offers a shared domain between MIS and RMIS, see Fig. 1.14. The shared domain is essential for the realization of the proposed IL pipeline, see Fig. 1.19, without which imitating human experts with a robot might be difficult. Many other works for laparoscopic camera motion automation in the image domain exist, and were reviewed in Section 1.5.2, yet most of them adhere to the tool following assumption, which we evidently rejected therein. It is, however, a priori not obvious how else one could formulate a visual servo instead. In this chapter, we argue that previous works are missing the bigger picture. We take a step back and attempt to shift the focus from tools to organs by treating the visual servoing task as a registration problem.

Arguably, registration might not seem a good automation paradigm, since surgical scenes are dynamic and change over time, which is likely why no one attempted it. Therefore, it is important to understand the context within this work. We are ultimately not interested in global registration but registration of temporal changes from $\hat{s}_t$ to $\hat{s}_{t+1}$, i.e. desired image space actions $\hat{a}_t^*$, that is the human expert policy. Now, prior to predicting actions through IL, Chapter 4, and Chapter 5, thus deviating from learning auxiliary tasks, refer Section 1.5.3, we choose to prove that, indeed, the hypothesized actions, i.e. homographies, refer Section 1.6.2.3, may be executed under the RCM constraint. In this work, we contribute exactly that. We formulate an image-based visual servo for executing the embodiment-invariant action $\hat{a}_t^*$. Deviating from the dominant methods, the proposed image-based visual servo requires neither explicit tool and camera positions nor any explicit image depth information, whilst satisfying a RCM constraint on a serial arm manipulator, see Fig. 1.18, thereby obeying Section 1.3.3.2 - System Considerations.

Built on top of the image-based visual servo, we propose a semi-autonomous scheme, where actions $\hat{a}_t^*$ are generated by the user through selecting target views. The approach allows a user to build a graph of desired views, from which, once built, views can be manually chosen and automatically servoed to irrespective of robot-patient frame transformation changes. This scheme targets level three autonomy, refer Section 1.1, as we gradually pave the way towards level five autonomy in the remainder of this thesis.

3.1.1 Limitations of Current Approaches and Contributions

The majority of existing methods rely on the tool distance to infer a control law. Only in [Ma 2019; Ma 2020; Aghakhani 2013; Yang 2019; Li 2020a; Osa 2010], the

position of arbitrary points w.r.t. the camera frame is fed back to the robot. All of the existing methods rely on relative positions, which either requires tool and camera positions or depth images. Position data might only be accessible in a fully robotic setup and image depth is difficult to estimate in a dynamic surgical environment from a monocular camera. Stereoscopic images are usually not available in robot assisted surgery.

Our paper addresses the above limitations with the following contributions:

- We introduce a visual servo that navigates towards desired images rather than towards points.
- We introduce a collaborative semi-autonomous scheme where the surgeon selects desired views.
- We formulate a visual servo control law that depends neither on explicit tool and camera positions nor on depth information.

These are achieved with a programmable RCM, as it, in contrast to a mechanical RCM, is more flexible.

This paper is structured as follows. In Section 3.2, we introduce the necessary theoretical background and the derivation of the proposed visual servoing task. In Section 3.3, we explain implementation details and the robotic setup. Results are provided in Section 3.4, and conclusions in Section 3.5.

3.2 Methods

Here, we first introduce the composite Jacobian for control in Section 3.2.1. Then, we extend it by a novel homography-based task function in Section 3.2.2, and describe the processing pipeline in Section 3.2.3. In the following, scalars are depicted by lower case letters, vectors through bold lower case letters, and matrices as bold upper case letters. A point x is described with respect to frame F as ${}^{\mathrm{F}}\mathbf{x}$.

3.2.1 Task Control with Remote Center of Motion Objective

For the task control with RCM objective, we follow the derivation of Aghakhani *et al.* [Aghakhani 2013]. Therefore, as schematically shown in Fig. 3.1, an open kinematic chain is attached to reference frame W. An endoscope is attached to the chain. It originates at position ${}^{\mathrm{W}}\mathbf{x}_i$ and has its camera frame at position ${}^{\mathrm{W}}\mathbf{x}_{i+1}$. The endoscope enters the patient through the trocar at position ${}^{\mathrm{W}}\mathbf{x}_{\mathrm{trocar}}$. The RCM

position ${}^{\mathrm{W}}\mathbf{x}_{\mathrm{RCM}}$ is required to lie along the line connecting ${}^{\mathrm{W}}\mathbf{x}_i$ to ${}^{\mathrm{W}}\mathbf{x}_{i+1}$, hence

$$ {}^{\mathrm{W}}\mathbf{x}_{\mathrm{RCM}} = {}^{\mathrm{W}}\mathbf{x}_i + \lambda \left({}^{\mathrm{W}}\mathbf{x}_{i+1} - {}^{\mathrm{W}}\mathbf{x}_i \right), \tag{3.1} $$

where the scalar $\lambda \geq 0$ is proportional to the entry depth. $\lambda = 0$ corresponds to maximal insertion. The endoscope's translational velocity at position ${}^{\mathrm{W}}\mathbf{x}_{\mathrm{RCM}}$ has to remain zero for the endoscope to reside at the trocar ${}^{\mathrm{W}}\mathbf{x}_{\mathrm{trocar}}$. It was derived in [Aghakhani 2013] as

$$ {}^{\mathrm{W}}\dot{\mathbf{x}}_{\mathrm{RCM}} = \begin{bmatrix} \mathbf{J}_i^v + \lambda(\mathbf{J}_{i+1}^v - \mathbf{J}_i^v) \\ {}^{\mathrm{W}}\mathbf{x}_{i+1} - {}^{\mathrm{W}}\mathbf{x}_i \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \dot{\mathbf{q}} \\ \dot{\lambda} \end{bmatrix}, \tag{3.2} $$

where $\mathbf{J}_i^v$, $\mathbf{J}_{i+1}^v$ are the Jacobians' top three rows, therefore the translational parts, corresponding to points ${}^{\mathrm{W}}\mathbf{x}_i$, ${}^{\mathrm{W}}\mathbf{x}_{i+1}$ w.r.t. the world frame, $\dot{\mathbf{q}}$ are the instantaneous joint velocities, and $\dot{\lambda}$ is the rate of change of entry depth. (3.2) can be rewritten as

$$ {}^{\mathrm{W}}\dot{\mathbf{x}}_{\mathrm{RCM}} = \mathbf{J}_{\mathrm{RCM}} \begin{bmatrix} \dot{\mathbf{q}} \\ \dot{\lambda} \end{bmatrix}. \tag{3.3} $$

Expanding on [Aghakhani 2013], we introduce a feedback to λ by projecting the trocar position $\mathbf{x}_{\mathrm{trocar}}$ onto the endoscope via

$$ \lambda = \frac{({}^{\mathrm{W}}\mathbf{x}_{i+1} - {}^{\mathrm{W}}\mathbf{x}_i)^{\mathrm{T}}({}^{\mathrm{W}}\mathbf{x}_{\mathrm{trocar}} - {}^{\mathrm{W}}\mathbf{x}_i)}{||{}^{\mathrm{W}}\mathbf{x}_{i+1} - {}^{\mathrm{W}}\mathbf{x}_i||_2^2}. \tag{3.4} $$

(3.3) can be further extended by a task as follows

$$ \begin{bmatrix} \dot{\mathbf{t}} \\ {}^{\mathrm{W}}\dot{\mathbf{x}}_{\mathrm{RCM}} \end{bmatrix} = \begin{bmatrix} \mathbf{J}_{\mathrm{t}} & \mathbf{0}_{n_{\mathrm{t}} \times 1} \\ \mathbf{J}_{\mathrm{RCM}} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{q}} \\ \dot{\lambda} \end{bmatrix}, \tag{3.5} $$

where $\dot{\mathbf{t}}$ is the task velocity with task dimension n_{t} and $\mathbf{J}_{\mathrm{t}}$ is the task Jacobian. (3.5) can be turned into a PID controller

$$ \begin{bmatrix} \dot{\mathbf{q}} \\ \dot{\lambda} \end{bmatrix} = \mathbf{J}_{\mathrm{cp}}^{\dagger} \left(\mathbf{K}^{\mathrm{p}} \begin{bmatrix} \mathbf{e}_{\mathrm{t}}^{\mathrm{p}} \\ {}^{\mathrm{W}}\mathbf{e}_{\mathrm{RCM}}^{\mathrm{p}} \end{bmatrix} + \mathbf{K}^{\mathrm{i}} \begin{bmatrix} \mathbf{e}_{\mathrm{t}}^{\mathrm{i}} \\ {}^{\mathrm{W}}\mathbf{e}_{\mathrm{RCM}}^{\mathrm{i}} \end{bmatrix} + \mathbf{K}^{\mathrm{d}} \begin{bmatrix} \mathbf{e}_{\mathrm{t}}^{\mathrm{d}} \\ {}^{\mathrm{W}}\mathbf{e}_{\mathrm{RCM}}^{\mathrm{d}} \end{bmatrix} \right), \tag{3.6} $$

where $\mathbf{J}_{\mathrm{cp}}^{\dagger}$ is the pseudo-inverse of the composite Jacobian from ((3.5)), $\mathbf{e}_t^{\mathrm{p/i/d}}$ and ${}^{\mathrm{W}}\mathbf{e}_{\mathrm{RCM}}^{\mathrm{p/i/d}}$, are the proportional, integral, and differential errors for the task and the RCM, respectively, and $\mathbf{K}^{\mathrm{p/i/d}}$ are the diagonal gain matrices. Therein, ${}^{\mathrm{W}}\mathbf{e}_{\mathrm{RCM}}^{\mathrm{i/d}}$ are computed as the integral, and the differential of the proportional error ${}^{\mathrm{W}}\mathbf{e}_{\mathrm{RCM}}^{\mathrm{p}} = {}^{\mathrm{W}}\mathbf{x}_{\mathrm{trocar}} - {}^{\mathrm{W}}\mathbf{x}_{\mathrm{RCM}}$.

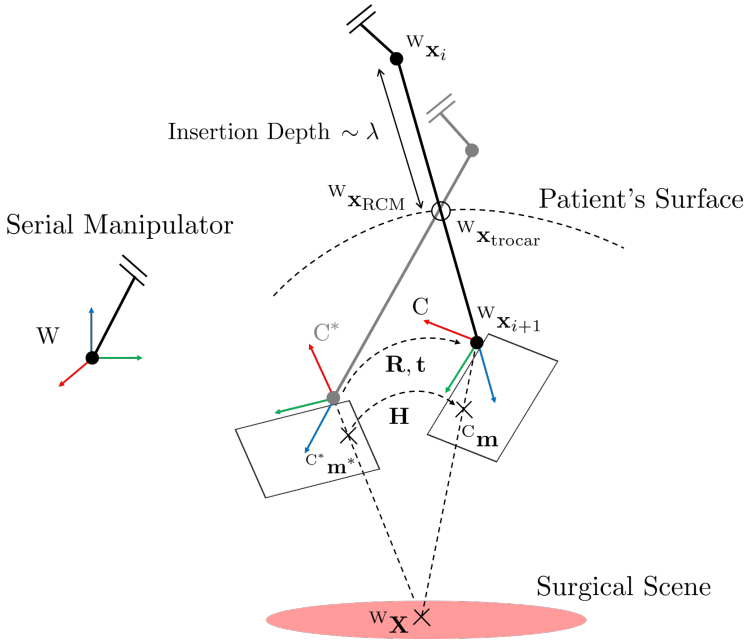


Figure 3.1: Schematic illustration of the setup: The axes' RGB coloring corresponds to XYZ, respectively. A serial manipulator is connected to the world frame W. The endoscope spans from ${}^{\mathrm{W}}\mathbf{x}_i$ to ${}^{\mathrm{W}}\mathbf{x}_{i+1}$ and it enters the trocar, which lies at $\mathbf{x}_{\mathrm{trocar}}$. The camera rotates around the RCM ${}^{\mathrm{W}}\mathbf{x}_{\mathrm{RCM}}$ and its entry depth is proportional to $\lambda \geq 0$. The camera observes the surgical scene (pink) from different frames C and C*.

In the following section, we introduce a homography-based visual servoing task.

3.2.2 Homography-based Visual Servoing Task

Suppose point ${}^{\mathrm{W}}\mathbf{X}$ is projected from a plane, i.e. the surgical scene, onto normalized coordinates $\mathbf{m}^*$ in camera frame C^*, see Fig. 3.1, via

$$ {}^{\mathrm{C}^*}\mathbf{m}^* = \frac{1}{{}^{\mathrm{C}^*}Z^*} \begin{bmatrix} {}^{\mathrm{C}^*}X^* & {}^{\mathrm{C}^*}Y^* & {}^{\mathrm{C}^*}Z^* \end{bmatrix}^{\mathrm{T}}, \tag{3.7} $$

which means it is observed by the camera as

$$ {}^{\mathrm{C}^*}\mathbf{p}^* = \mathbf{K}\,{}^{\mathrm{C}^*}\mathbf{m}^*, \tag{3.8} $$

in pixel coordinates ${}^{\mathrm{C}^*}\mathbf{p}^* = \begin{bmatrix} u^* & v^* & 1 \end{bmatrix}^{\mathrm{T}}$, with the camera's intrinsic parameters $\mathbf{K}$. Should the camera move under rotation $\mathbf{R}$ and translation $\mathbf{t}$, the points in normalized coordinates will change according to a homography $\mathbf{H}$ such that [Benhimane 2006]

$$ \frac{{}^{\mathrm{C}}Z}{{}^{\mathrm{C}^*}Z^*}\,{}^{\mathrm{C}}\mathbf{m} = \mathbf{H}\,{}^{\mathrm{C}^*}\mathbf{m}^* \tag{3.9} $$

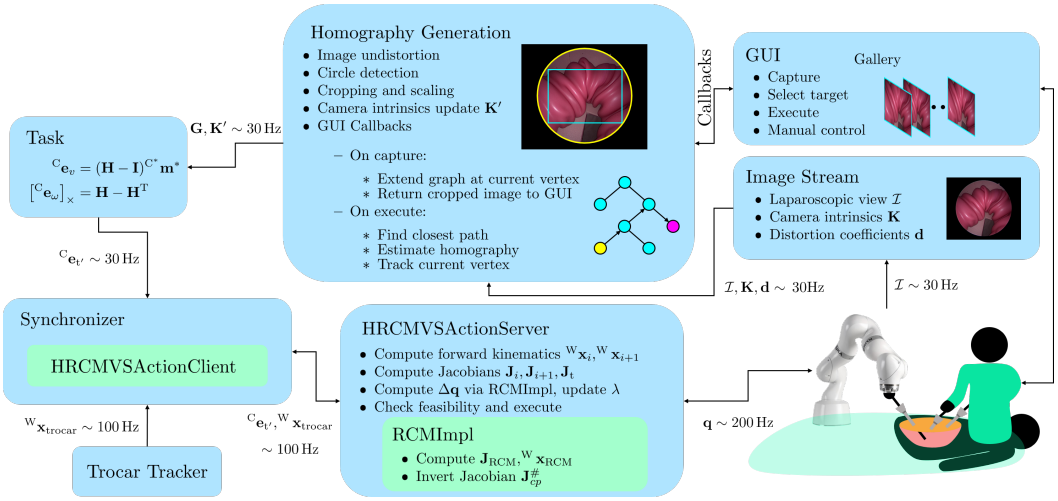


Figure 3.2: Processing pipeline. A surgeon manually controls the robot through a GUI, collecting desired views along the way. The images are pre-processed, and a graph of desired views is built in the background by the homography generation node. Once built, the surgeon selects desired views through the GUI, which triggers a shortest path finding from the current vertex (yellow), to the desired one (pink), and the execution of subsequent homography estimations that lead to the target.

In pixel coordinates this can be written as

$$\frac{{}^{\mathrm{C}}Z}{{}^{\mathrm{C}^*}Z^*}\,{}^{\mathrm{C}}\mathbf{p} = \mathbf{G}\,{}^{\mathrm{C}^*}\mathbf{p}^*, \tag{3.10}$$

with the projective homography $\mathbf{G}$, for which the following relation holds

$$\mathbf{H} = \mathbf{K}^{-1}\mathbf{G}\mathbf{K}. \tag{3.11}$$

As shown in [Benhimane 2006], the task error ${}^{\mathrm{C}}\mathbf{e}_{\mathrm{t}'} = \begin{bmatrix} {}^{\mathrm{C}}\mathbf{e}_v & {}^{\mathrm{C}}\mathbf{e}_\omega \end{bmatrix}^{\mathrm{T}}$ that urges to minimize the distance between the desired projection of ${}^{\mathrm{W}}\mathbf{X}$, ${}^{\mathrm{C}^*}\mathbf{m}^*$, and the current one ${}^{\mathrm{C}}\mathbf{m}$, can be obtained purely from the homography that relates those points in normalized coordinates via

$$\begin{aligned} {}^{\mathrm{C}}\mathbf{e}_v &= (\mathbf{H} - \mathbf{I})\,{}^{\mathrm{C}^*}\mathbf{m}^* \\ \left[{}^{\mathrm{C}}\mathbf{e}_\omega\right]_\times &= \mathbf{H} - \mathbf{H}^{\mathrm{T}}, \end{aligned} \tag{3.12}$$

where $\left[{}^{\mathrm{C}}\mathbf{e}_\omega\right]_\times$ is the skew symmetric matrix of ${}^{\mathrm{C}}\mathbf{e}_\omega$. The task error ${}^{\mathrm{C}}\mathbf{e}_{\mathrm{t}'}$ is described in body coordinates. It can be transferred to the world frame W through rotation,

which is proportional to camera frame's instantaneous velocity

$$\begin{bmatrix} {}^{\mathrm{W}}\mathbf{R}_{\mathrm{C}} & \mathbf{0} \\ \mathbf{0} & {}^{\mathrm{W}}\mathbf{R}_{\mathrm{C}} \end{bmatrix} {}^{\mathrm{C}}\mathbf{e}_{\mathrm{t}'} = {}^{\mathrm{W}}\,\mathbf{e}_{\mathrm{t}'} \sim \mathbf{J}_{i+1}\dot{\mathbf{q}} \tag{3.13}$$

where ${}^{\mathrm{W}}\mathbf{R}_{\mathrm{C}}$ is the rotation of the camera frame with respect to the world frame, and $\mathbf{J}_{i+1}$ is the camera frame's Jacobian, including its rotational contributions. Only 4 DoF can be controlled at a time after imposing the RCM, which constraints 2 DoF. To capture this, we introduce operator $\mathbf{P}$ that projects the camera frame body velocity onto the remaining DoF. Together with ((3.13)), this yields

$$\mathbf{P}_{a/b}\,{}^{\mathrm{C}}\mathbf{e}_{\mathrm{t}'} = {}^{\mathrm{C}}\,\mathbf{e}_{\mathrm{t}_{a/b}} \sim \mathbf{P}_{a/b} \begin{bmatrix} {}^{\mathrm{C}}\mathbf{R}_{\mathrm{W}} & \mathbf{0} \\ \mathbf{0} & {}^{\mathrm{C}}\mathbf{R}_{\mathrm{W}} \end{bmatrix} \mathbf{J}_{i+1}\dot{\mathbf{q}}. \tag{3.14}$$

The projection operator $\mathbf{P}_{a/b}$ can take different forms, such that the task error is mapped onto any of the decoupled remaining DoF via

$$\mathbf{P}_a = \begin{bmatrix} \mathbf{I}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{1\times3} & 0 \quad 0 \quad 1 \end{bmatrix}, \mathbf{P}_b = \begin{bmatrix} 0 \quad 0 \quad 1 & \mathbf{0}_{1\times3} \\ \mathbf{0}_{3\times3} & \mathbf{I}_{3\times3} \end{bmatrix}. \tag{3.15}$$

Therefore, $\mathbf{P}_a$ maps the task error ${}^{\mathrm{C}}\mathbf{e}_{\mathrm{t}'}$ to its translational parts and the rotation about the optical axis, and $\mathbf{P}_b$ maps it to its rotational part and the error along the optical axis. We identify the case sensitive contributions of ((3.14)) as the task Jacobian from ((3.5)) and the task error from ((3.6)), which yields

$$\mathbf{J}_{\mathrm{t}} = \mathbf{P}_{a/b} \begin{bmatrix} {}^{\mathrm{C}}\mathbf{R}_{\mathrm{W}} & \mathbf{0} \\ \mathbf{0} & {}^{\mathrm{C}}\mathbf{R}_{\mathrm{W}} \end{bmatrix} \mathbf{J}_{\mathrm{i+1}},\ \mathbf{e}_{\mathrm{t}}^{\mathrm{p}} = \begin{cases} {}^{\mathrm{C}}\mathbf{e}_{\mathrm{t}_a} = \begin{bmatrix} {}^{\mathrm{C}}\mathbf{e}_v & {}^{\mathrm{C}}e_{\omega_z} \end{bmatrix}^T \\ {}^{\mathrm{C}}\mathbf{e}_{\mathrm{t}_b} = \begin{bmatrix} {}^{\mathrm{C}}e_{v_z} & {}^{\mathrm{C}}\mathbf{e}_{\omega} \end{bmatrix}^T \end{cases} \tag{3.16}$$

This results in a task dimension $n_{\mathrm{t}} = 4$, which means that together with the RCM objective that introduces 3 constraints and adds the additional DoF λ, the robot has to have at least 6 DoF.

3.2.3 Processing Pipeline

An overview of the processing pipeline is depicted in Fig. 3.2. A surgeon first controls the endoscope from within the camera's reference frame via the keyboard. Images of desired views are manually taken along the way and are used to construct a graph, wherein each vertex is an image. This is done within the *homography generation* node.

Initially, camera calibration considering an underlying radial/tangential distortion

model is carried out to obtain the distortion coefficients and the camera intrinsics. Following that, an eye in hand calibration is performed to locate the camera frame position ${}^{\mathrm{W}}\mathbf{x}_{i+1}$, and ${}^{\mathrm{W}}\mathbf{x}_i$ is set to lie along the negative optical axis at the endoscope's length, see Fig. 3.1.

Each image $\mathcal{I}$ that is processed within the *homography generation* node undergoes distortion removal, followed by an intensity-based automatic detection of the endoscopic boundary circle. Therein, the image is smoothed with a bilateral filter and thresholded in HSV image space to obtain a binary mask. The circle's center is computed as the center of mass, and its radius is obtained from the steepest gradient of the marginalized binary mask. If the illumination in the endoscopic view is below a certain value, then the last known center and radius are considered instead. The maximum rectangle of a given aspect ratio that fits into the extracted circle is then cropped from the image $\mathcal{I}$. The crop is further rescaled. The camera intrinsics are updated accordingly from $\mathbf{K}$ to $\mathbf{K}'$ by offsetting and scaling the principal point.

Once the graph is built, the surgeon can browse through the image gallery, as shown in Fig. 3.2, where each image corresponds to a vertex within the graph. The surgeon may then select a desired view and execute the visual servo. This will trigger a Dijkstra search for the closest path from the current vertex to the desired view/vertex at constant cost per edge. This path is executed sequentially. Therefore, the homography $\mathbf{G}$ from the next vertex to the current view is computed for the visual servo. To compute the homography, we extract image features and their descriptors with a SURF feature detector [Bay 2006]. For each feature in the target view, the two nearest neighbors are found in the current view, and, via Lowe's ratio test [Lowe 2004], only features with distinctive descriptors are kept. The homography that maps features from the target view to the current view is then determined under RANSAC outlier rejection.

The updated camera intrinsics $\mathbf{K}'$, together with the desired homography $\mathbf{G}$, are then sent down the pipeline to first transform the homography from pixel coordinates to normalized coordinates via ((3.11)) and then to compute the desired task ${}^{\mathrm{C}}\mathbf{e}_{\mathrm{t}'}$ from ((3.12)). The update rate of these operations are restricted by the camera frame rate, which is why the desired trocar position ${}^{\mathrm{W}}\mathbf{x}_{\mathrm{trocar}}$ is sent separately to the synchronizer node, see Fig. 3.2. The synchronizer node takes a homography RCM visual servo action client, *HRCMVSActionClient*, which request the *HRCMVSActionServer* to execute the desired task ${}^{\mathrm{C}}\mathbf{e}_{\mathrm{t}'}$, while maintaining a desired trocar position ${}^{\mathrm{W}}\mathbf{x}_{\mathrm{trocar}}$.

The *HRCMVSActionServer* implements a state machine, which rejects infeasible

requests. It computes the forward kinematics as well as the Jacobians and computes a joint position update $\Delta\mathbf{q} = \Delta t\dot{\mathbf{q}}$ via ((3.6)) in the RCM implementation *RCMImpl*, where Δt is the control interval. The desired joint positions are then sent to the robot.

3.3 Experimental Setup

This section gives an overview of the robotic system and its components in Section 3.3.1. Following that, clinically relevant questions and the evaluation protocol are addressed in Section 3.3.2.

3.3.1 Robotic System

Our experimental setup, see Fig. 1.18, uses a KUKA LBR Med 7 R800 robot. To control it, we created a bridge to ROS by wrapping the FRI [Schreiber 2010] with ROS' Hardware Interface functionality. We use a Storz Endocameleon Hopkins Telescope, from which we capture images using a Storz TH 102 H3-Z FI camera head. The endoscope is mounted to the LBR Med 7 R800 robot with a custom designed 3D print. For illumination, we connect a Storz TL 300 Power LED 300 light source to the endoscope. The image feed is output to SDI, which we convert to HDMI with a Monoprice 3G SDI to HDMI converter. We then grab the HDMI signal with a DeckLink 4K Extreme 12G and stream it onto the ROS network.

3.3.2 Clinical Scenario Evaluation Protocol

The proposed method is evaluated in the laparoscopic setup shown in Fig. 1.18. We utilize a Szabo Pelvic Trainer to simulate a trocar. A Kyoto Kagaku colon rectum tube is inserted into the Szabo Pelvic Trainer to model a laparoscopic view of the abdomen. The clinical procedure is then modeled as follows. The robot initially drives the endoscope to the trocar and λ in ((3.1)) is set to 1. Following that, the user mounts the camera and the light source to the laparoscope. The user then drives the laparoscope through the trocar into the phantom.

In the phantom, we identify four clinically relevant views of the scene. These views include an overview of the scene, a view of the tool insertion area towards the abdominal wall, and two close-ups, one for further examination. For visual servoing between these views in a clinical scenario, these three objectives are of importance

- Servoing from any current to any target view.
- Servoing to target views under tool motion.

- Servoing to target views after phantom repositioning.

To address these scenarios, we design three experiments. For all experiments, after the laparoscope insertion, the user moves to the overview of the surgical scene, where the first image is taken through the GUI, which corresponds to the graph's root view/vertex, see Fig. 3.2. We measure the deviation of the RCM from the trocar position, record the MPD of SURF features from the current to the desired view, the task error, execution time, joint angles, and the camera tip position.

3.3.2.1 Servoing from any current view to any target view

In this scenario we investigate the system's capability to autonomously execute extreme view changes. The user moves from the overview to a close-up, from where the scene is further examined. The laparoscope is then moved manually to grant view of the tool insertion area. At this stage, tools would be inserted into the patient and the user would begin to operate. Therefore, the user selects the close-up view through the GUI and executes the autonomous visual servo towards it.

3.3.2.2 Servoing to target views under tool motion

In this scenario we investigate autonomous visual servoing towards desired views under tool motion. Therefore, the user moves the laparoscope from the overview to the tool insertion area. Tools are then inserted and the user is asked to perform a sample task, which involves moving small LASTT Training Package rings. The visual servo simultaneously navigates back towards the overview.

3.3.2.3 Servoing to target views after phantom repositioning

In this scenario we investigate the system's invariance under patient motion. Therefore, we reposition the phantom and execute the visual servo to autonomously readjust the overview. We include both phantom rotation and tilting.

3.4 Results

In this section, we first present generic findings in Section 3.4.1, followed by quantitative measurements for the evaluation protocol from Section 3.3.2, in Section 3.4.2.

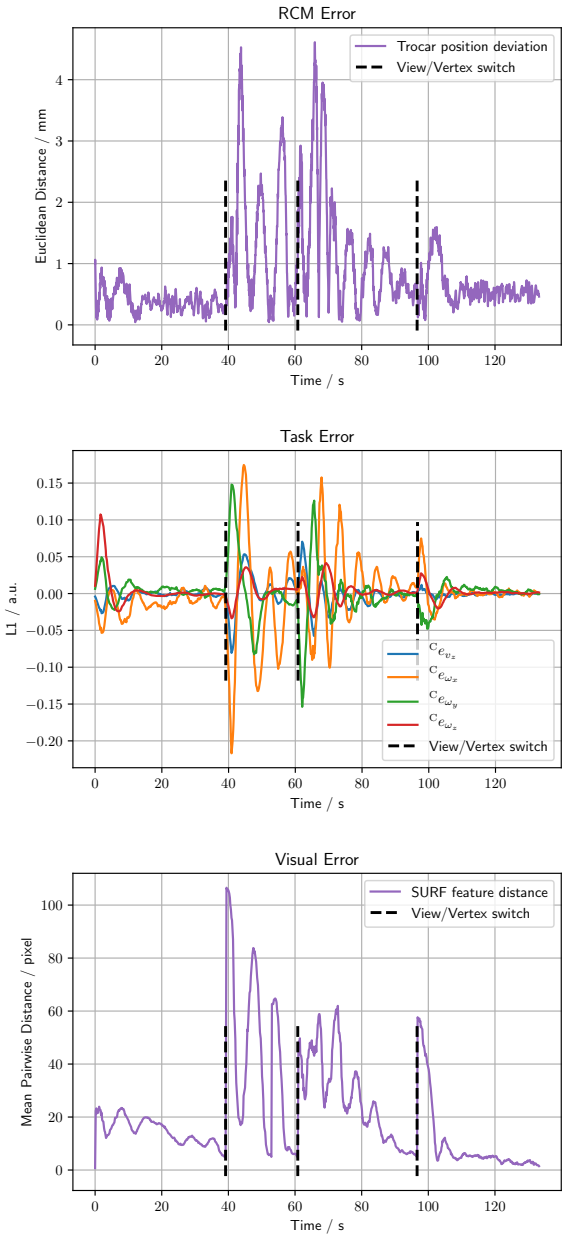


Figure 3.3: RCM deviation (top) and task error evolution (bottom) over time for the protocol in Section 3.3.2.1. The visual servo autonomously servos from the tool insertion area to the close-up. Target views/vertices are updated along the way, as indicated by the black dotted lines. The RCM error of 1 mm or less compares well to literature, however, the errors of 1 mm and above at the transition points are larger than reported in compared work and further gain fine-tuning might be necessary.

3.4.1 Generic Results

In practice we found that controlling the camera frame's rotational DoF, using $\mathbf{P}_b$ in ((3.15)), leads to more stable solutions. We tried to invert the task part of the composite Jacobian from ((3.6)) within the Nullspace of the RCM Jacobian, but obtained more flexible solutions by computing the pseudo-inverse as a damped least squares solution from the SVD with a damping factor of 5e−4. Through hand tuning, we got good results with the following gain matrices

$$\begin{aligned}\mathbf{K}^{\mathrm{p}} &= \mathrm{diag}(1.2, 1.5, 1.5, 1.8, 1\mathrm{e}2, 1\mathrm{e}2, 1\mathrm{e}2)\\ \mathbf{K}^{\mathrm{i}} &= \mathrm{diag}(3\mathrm{e}{-3}, 2.5\mathrm{e}{-3}, 2.5\mathrm{e}{-3}, 1.5\mathrm{e}{-3}, 0, 0, 0)\\ \mathbf{K}^{\mathrm{d}} &= \mathrm{diag}(6\mathrm{e}{-2}, 5\mathrm{e}{-2}, 5\mathrm{e}{-2}, 3\mathrm{e}{-2}, 0, 0, 0).\end{aligned}$$

The integral term therein helped remove a steady state error in the homography-based image alignments. The desired homography extraction proved noisy but correct on average, so we introduced a moving average filter on the task error ${}^{\mathrm{C}}\mathbf{e}_{\mathrm{t}}$ with a buffer length of 10 at a frame rate of 30 fps. The sequential execution of desired views was greatly sped up by calling early convergence for intermediate vertices/views at a MPD of 5 pixels and a final convergence at a MPD of 1.5 pixels.

3.4.2 Clinical Scenario Results

3.4.2.1 Servoing from any current view to any target view

In this section we investigate the trajectory from tool insertion view to close-up, see Section 3.3.2.1. The task error and the RCM deviation from the trocar position are depicted in Fig. 3.3. It can be seen that the deviation from the trocar position stays below 4.6 mm, at an average deviation of 0.8 ± 0.8 mm. The task error converges for all vertices/views. The final task error corresponds to a camera tip deviation of 0.4 mm, when compared to the desired position. The joint angles deviate on average by $8.2 \pm 6.0^\circ$ from the initial configuration.

3.4.2.2 Servoing to target views under tool motion

For this measurement, the visual servo navigates from the tool insertion area to the overview under tool motion, see Section 3.3.2.2. The trajectory with all intermediate and the final vertex/view is shown in Fig. 3.4. It can be seen that, despite tool motion, the visual servo converges at pixel accuracy towards the desired views. The final camera position deviates by 1.4 mm to the desired one. The robot joint angles deviate on average by $1.1 \pm 1.1^\circ$ from the initial configuration. A video of

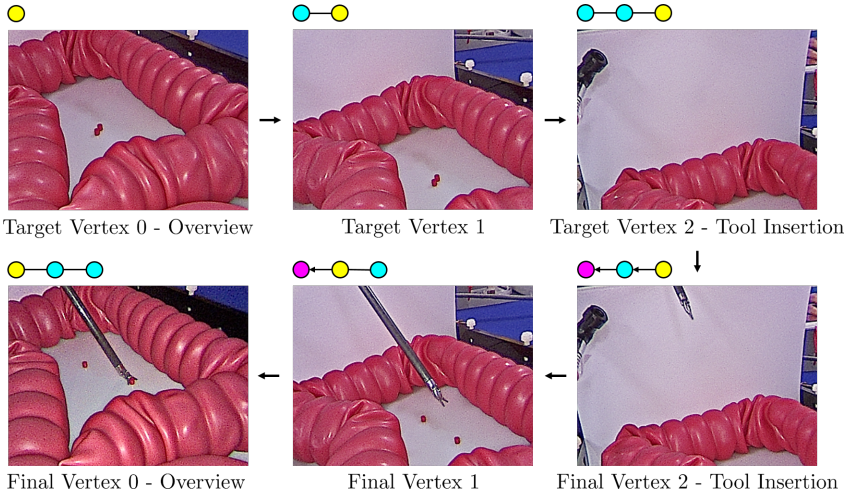


Figure 3.4: Servoing under tool motion, see Section 3.3.2.2. Initially, the graph is built in manual control mode (top row), yellow indicates the current vertex. The visual servo is then executed to navigate back from the tool insertion to the overview (bottom row). Pink indicates the target vertex.

this experiment is provided under [1].

3.4.2.3 Servoing to target views after phantom repositioning

In this section we investigate the convergence of the visual error after phantom repositioning, see Section 3.3.2.3. We perform clockwise and counterclockwise repositioning as well as phantom tilting. We keep the trocar at the initial position. The camera frame then rotates and translates towards a position that minimizes the visual error. The translation $\Delta\mathbf{x}_{i+1}$ and the angle axis rotation angle α are listed in Table 3.1. It can be seen that the robotic laparoscope performs significant motion to readjust the view. The MPD is minimized to pixel range and the final deviation from the trocar remains in the submillimeter scale for all cases.

3.5 Conclusion and Future Work

In this work we introduced a visual servo that is independent of depth information and explicit tool and camera positions. The introduced method simultaneously respects a programmable RCM. Our method was successfully integrated into a robotic setup and clinically relevant scenarios were investigated on an abdominal phantom.

[1] `https://drive.google.com/file/d/1UCr__R2_7xit6TTq3T9pTIEg1fMsfT3j/view?usp=sharing`

Table 3.1: CW, and CCW repositioning, and phantom tilting, corresponding to the protocol in Section 3.3.2.3. $\Delta\mathbf{x}_{i+1}$ indicates the camera motion, α the angle axis rotation angle from initial to final camera rotation, $\Delta\mathbf{q}$ the joint angle position change, $\mathbf{e}_{\text{RCM}}$ the final deviation of the RCM from the trocar, and MPD the final visual error.

Metric	CW	CCW	Tilt
$\Delta\mathbf{x}_{i+1}$ / mm	10.4	6.7	4.7
α / °	16.6	10.2	4.8
$\Delta\mathbf{q}$ / °	20.5 ± 12.0	17.4 ± 13.4	2.6 ± 2.3
${}^{\text{W}}\mathbf{e}^{\text{p}}_{\text{RCM}}$ / mm	0.1	0.2	0.07
MPD / pixel	3.2 ± 2.5	2.0 ± 1.0	1.4 ± 1.2

Results Discussion It was shown in Section 3.4.2.1 that the proposed composite Jacobian PID controller with homography-based task simultaneously minimizes the RCM and the visual servo objective. The integral term proved helpful to remove a steady state error in the image alignment. The homography estimation was noisy due to feature sparseness and required for average filtering. The graph representation allowed for visual servoing between images that were not relatable by a single homography transformation. In Section 3.4.2.2, tools were successfully introduced into the scene. It is to be noted that the tools were initially not present in the target views, which removed potential image misalignment. In Section 3.4.2.3 the phantom was repositioned significantly with a constant trocar position and image readjustment was successfully demonstrated. The MPD got close to perfect alignment, however, the trocar was possibly moved slightly during repositioning, which made perfect convergence not possible. The robot's joint angles did not always return to their initial configuration. The camera position converged in submillimeter range to its target.

Future Work We successfully demonstrated that our visual servo navigates the camera in submillimeter range without depth information or explicit tool and camera positions. This proves the future potential for safe patient application and it circumvents time-consuming registration procedures. As our setup has one redundant DoF, the robot did not always return to its initial configuration. This might be handled by introducing joint state objectives to the Jacobian's nullspace. While our visual servo is independent of registration procedures, the RCM requires initialization, and tracking. In future work, the controller might be updated as to incorporate force-torque sensing to update the RCM. Although the environment was mostly static, the homography estimation was noisy, and image augmentations could have been introduced to improve realism. In future research, one might,

therefore, incorporate homography estimation that is invariant under object motion and robust under feature sparseness, using deep learning approaches, which we will introduce in the following Chapter 4.

Chapter 4

Laparoscopic Camera Motion Extraction from Dynamic Surgical Scenes

Table of Contents

Disclaimer This Chapter 4 is an *in extenso* reproduction of [Huber 2022]. Only Section 4.1 was altered to highlight additional context within the scope of this thesis.

4.1 Introduction

Having successfully established a homography-based visual servo under RCM constraint in the previous Chapter 3, e.g. Fig. 3.4, which is based on simple optimal control and, unlike end-to-end learned robot policies, inherently safe for clinical use, this chapter will now investigate the generation of large scale state-action pairs $(\hat{s}_t, \hat{a}_t^*)$. The generation of these state-action pairs is a prerequisite for extracting the human expert policy through IL methods, refer Section 1.5.5, which will be delved further into in Chapter 5. For now, the focus will solely lie on extracting actions to form state-action pairs.

The slow transition of IL methods to laparoscopic camera motion automation can, indeed, be attributed to the lack of state-action pairs, as explained in Section 1.6.2.1. This lack has historically lead to the emergence of rule-based approaches, see Section 1.5.2. In this chapter, we argue that is possible to extract actions, i.e. camera motion, from videos of laparoscopic interventions, using a DNN. This, however, is not a trivial endeavor, and hence other research vastly focuses on solving auxiliary vision tasks instead, as explained in Section 1.5.3. Therefore, clinical data of laparoscopic surgeries, see Table 1.3, remains unusable for IL. Consequentially, state of the art (SOTA) IL attempts rely on artificially acquired data, as was reviewed in Section 1.6.2.2.

The main difficulty for extracting camera motion from videos of laparoscopic interventions is to distinguish it from tool and object motion. In Section 1.6.2.2, we hypothesized that this might be possible through a supervised training procedure in which a DNN is regressed on isolating camera from object and tool motion. We suggested to exploit the clamping mechanism of the da Vinci® robot for this purpose, see Fig. 1.20. Therefore, hypothesizing one may train DNNs to such robustness they transfer from videos of RMIS to MIS for the purpose of camera motion extraction. The camera motion formulation, therein, will align with actions introduced in the previous Chapter 3, thus enabling the feedback loop, as shown in Fig. 1.19. The registration problem will thus be formulated via homographies, Section 1.6.2.3, (1.13). Interestingly, this line of research has applications in multiple other domains where registration despite organ motion is of relevance.

4.1.1 Contributions

In this work, we aim to extract camera motion from videos of laparoscopic interventions, thereby creating state-action-pairs for IL. To this end, we introduce a method that isolates camera motion (actions) from object and tool motion by solely relying

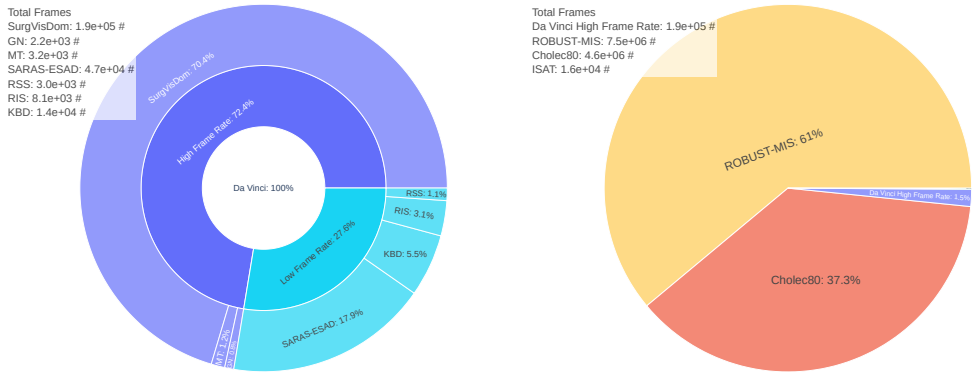
on observed images (states). To this end, DNNs are supervisedly trained to estimate camera motion while disregarding object, and tool motion. This is achieved by synthetically adding camera motion via a novel *homography generation algorithm* to a newly acquired dataset of camera motion free da Vinci® surgery image sequences. In this way, object, and tool motion reside within the image sequences, and the synthetically added camera motion can be regarded as the only source, and therefore ground truth, for camera motion estimation. Extensive experiments are carried out to identify modern network architectures that perform best at camera motion estimation. The DNNs that are trained in this manner are found to generalize well across domains, in that they transfer to vast laparoscopic datasets. They are further found to outperform classical camera motion estimators.

4.2 Related Work

Supervised deep homography estimation was first introduced in [DeTone 2016] and got improved through a hierarchical homography estimation in [Nowruzi 2017]. It got adopted in the medical field in [Bano 2020]. All three approaches generate a limited set of homographies, only train on static images, and use non-SOTA VGG-based network architectures [Simonyan 2015].

Unsupervised deep homography estimation has the advantage to be applicable to unlabelled data, e.g. videos. It was first introduced in [Nguyen 2018], and got applied to endoscopy in [Gomes 2019]. The loss in image space, however, cannot account for object motion, and only static scenes are considered in their works. Consequentially, recent work seeks to isolate object motion from camera motion through unsupervised incentives. Closest to our work is [Le 2020], where the authors generate a dataset of camera motion free image sequences. However, due to tool, and object motion, their data generation method is not applicable to laparoscopic videos, since it relies on motion free image borders. The authors in [Zhang 2020], provide the first work that does not need a synthetically generated dataset. Their method works fully unsupervised, but constraining what the network minimizes, is difficult to achieve.

Only [Le 2020] and [Zhang 2020] train DNNs on object motion invariant homography estimation. Contrary to their works, we train DNNs supervisedly. We do so by applying the data generation of [DeTone 2016] to image sequences rather than single images. We further improve their method by introducing a novel *homography generation algorithm* that allows to continuously generate synthetic homographies at runtime, and by using SOTA DNNs.



(a) da Vinci® surgery datasets. Included are: SurgVisDom [Zia 2021], GN [Giannarou 2013], MT [Mountney 2010], SARAS-ESAD [Bawa 2020], KBD [Hattab 2020], RIS [Allan 2019], RSS [Allan 2020]

(b) Laparoscopic datasets and HFR da Vinci® dataset from (a) for comparison. Included are: ROBUST-MIS [Maier-Hein 2021], Cholec80 [Twinanda 2017], ISAT [Bodenstedt 2018].

Figure 4.1: da Vinci® surgery and laparoscopic surgery datasets. Shown are relative sizes and the absolute number of frames. da Vinci® surgery datasets are often released at a low frame rate of 1 fps for segmentation tasks (a). Much more laparoscopic surgery data is available (b).

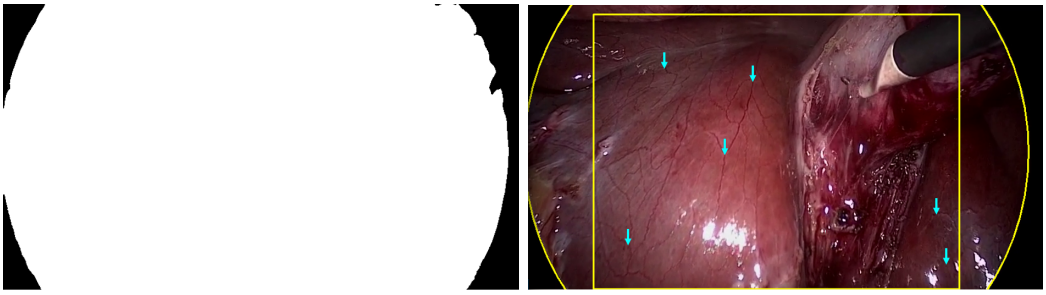
4.3 Materials and Methods

4.3.1 Data Preparation

Similar to [Le 2020], we initially find camera motion free image sequences, and synthetically add camera motion to them. In our work, we isolate camera motion free image sequences from da Vinci® surgeries, and learn homography estimation supervisedly. We acquire publicly available laparoscopic, and da Vinci® surgery videos. An overview of all datasets is shown in Fig. 4.1. Excluded are synthetic, and publicly unavailable datasets. da Vinci® surgery datasets, and laparoscopic surgery datasets require different pre-processing steps, which are described below.

4.3.1.1 da Vinci® Surgery Data Pre-Processing

Many of the da Vinci® surgery datasets are designed for tool or tissue segmentation tasks, therefore, they are published at a frame rate of 1 fps, see Fig. 4.1a. We merge all HFR datasets into a single dataset and manually remove image sequences with camera motion, which amount to 5% of all HFR data. We crop the remaining data to remove status indicators, and scale the images to 306×408 pixels, later to be cropped by the *homography generation algorithm* to a resolution of 240×320.



(a) Binary segmentation mask, obtained through thresholding the bilateral filtered image.

(b) Circular boundary detection and static landmarks (blue arrows). Landmarks are, e.g. identified at vein bifurcations.

Figure 4.2: Cholec80 dataset pre-processing, referring to Section 4.3.1.2. The black boundary circle is automatically detected. Landmarks are manually annotated and tracked over time (b).

4.3.1.2 Laparoscopic Surgery Data Pre-Processing

Laparoscopic images are typically observed through a Hopkins telescope, which causes a black circular boundary in the view, see Fig. 4.2. This boundary does not exist in da Vinci® surgery recordings. For inference on the laparoscopic surgery image sequences, the most straightforward approach is to crop the view. To this purpose, we determine the center and radius of the circular boundary, which is only partially visible. We detect it by randomly sampling N points $\mathbf{p}_i = (u_i, v_i)^{\mathrm{T}}$ on the boundary. This is similar to work in [Münzer 2013] and our later contributed graphics processing unit (GPU) accelerated and differentiable version in [Budd 2022] (accessible on GitHub[1]), but instead of computing an analytical solution, we fit a circle by means of a least squares solution through inversion of

$$\begin{bmatrix} 2u_0 & 2v_0 & 1 \\ & \vdots & \\ 2u_{N-1} & 2v_{N-1} & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} u_0^2 + v_0^2 \\ \vdots \\ u_{N-1}^2 + v_{N-1}^2 \end{bmatrix}, \tag{4.1}$$

where the circle's center is (x_0, x_1), and its radius is $\sqrt{x_2 + x_0^2 + x_1^2}$. We then crop the view centrally around the circle's center, and scale it to a resolution of 240×320. An implementation is provided on GitHub[2].

[1] `https://github.com/charliebudd/torch-content-area`
[2] `https://github.com/RViMLab/endoscopy`

4.3.1.3 Ground Truth Generation

One can simply use the synthetically generated camera motion as ground truth at train time. For inference on the laparoscopic dataset, this is not possible. We therefore generate ground truth data by randomly sampling 50 image sequences with 10 frames each from the Cholec80 dataset. In these image sequences, we find characteristic landmarks that are neither subject to tool, nor to object motion, see Fig. 4.2b. Tracking of these landmarks over time allows one to estimate the camera motion in between consecutive frames through (1.12).

4.3.2 Deep Homography Estimation

In this work we exploit the static camera in da Vinci® surgeries, which allows us to isolate camera motion free image sequences. The processing pipeline is shown in Fig. 4.3.

Image pairs are sampled from image sequences of the HFR da Vinci® surgery dataset of Fig. 4.1a. An image pair consists of an anchor image $\mathcal{I}_n$, and an offset image $\mathcal{I}_{n+t}$. The offset image is sampled uniformly from and interval $t \in [-T, T]$ around the anchor. The HFR da Vinci® surgery dataset is relatively small, compared to the laparoscopic datasets, see Fig. 4.1b. Therefore, we apply image augmentations to the sampled image pairs. They include transform to grayscale, horizontal, and vertical flipping, cropping, change in brightness, and contrast, Gaussian blur, fog simulation, and random combinations of those. Camera motion is then added synthetically to the augmented image $\mathcal{I}^{\text{aug}}_{n+t}$ via the *homography generation algorithm* from Section 4.3.3. A DNN, with a backbone, then learns to predict the homography $\mathbf{G}_{\text{4point}}$ between the augmented image, and the augmented image with synthetic camera motion at time step $n + t$.

4.3.3 Homography Generation Algorithm

In its core, the *homography generation algorithm* is based on the works of [DeTone 2016]. However, where [DeTone 2016] crop the image with a safety margin, our method allows to sample image crops across the entire image. Additionally, our method computes feasible homographies at runtime. This allows us to continuously generate synthetic camera motion, rather then training on a fixed set of precomputed homographies. The *homography generation algorithm* is summarized in Algorithm 2, and visualized in Fig. 4.3.

Initially, a *crop polygon* $\mathbb{P}_c$ is generated for the augmented image $\mathcal{I}^{\text{aug}}_n$. The *crop polygon* is defined through a set of points in the augmented image $\mathbb{P}_c = \{\mathbf{p}^c_i, i \in$

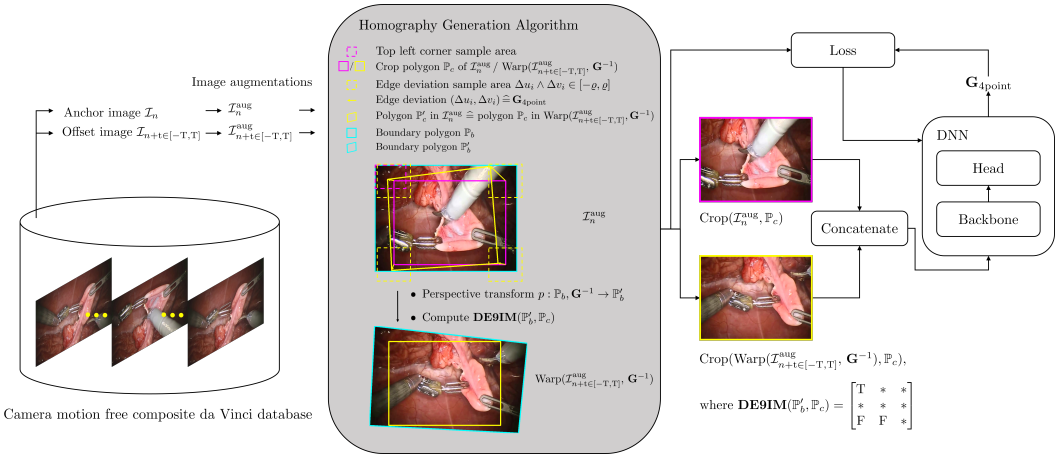


Figure 4.3: Deep homography estimation training pipeline. Image pairs are sampled from the HFR da Vinci® surgery dataset. The *homography generation algorithm* then adds synthetic camera motion to the augmented images, which is regressed through a backbone DNN.

$[0, 3]\}$, which span a rectangle. The top left corner $\mathbf{p}_0^c$ is randomly sampled such that the *crop polygon* $\mathbb{P}_c$ resides within the image border polygon $\mathbb{P}_b$, hence $\mathbf{p}_0^c \in ([0, h_b - h_c], [0, w_b - w_c])$, where h, and w are the height and width of the *crop*, and the *border polygon*, respectively. Following that, a random four point homography $\mathbf{G}_{4\text{point}}$ (1.13) is generated by sampling edge deviations $\Delta u_i \wedge \Delta v_i \in [-\varrho, \varrho]$. The corresponding inverse homography $\mathbf{G}^{-1}$ is used to warp each point of the border polygon $\mathbb{P}_b$ to $\mathbb{P}'_b$. Finally, the Dimensionally Extended 9-Intersection Model [Clementini 1994] is used to determine whether the warped polygon $\mathbb{P}'_b$ contains $\mathbb{P}_c$, for which we utilize the Python library *Shapely*[3]. If the thus found intersection matrix **DE9IM** satisfies

$$\mathbf{DE9IM}(\mathbb{P}'_b, \mathbb{P}_c) = \begin{bmatrix} T & * & * \\ * & * & * \\ F & F & * \end{bmatrix}, \tag{4.2}$$

the homography $\mathbf{G}^{-1}$ is returned, otherwise a new four point homograpy $\mathbf{G}_{4\text{point}}$ is sampled. Therein, $*$ indicates that the intersection matrix may hold any value, and T, F indicate that the intersection matrix must be true or false at the respective position. In the unlikely case that no homography is found after *maximum rollouts*, the identity $\mathbf{G}_{4\text{point}} = \mathbf{0}$ is returned. Once a suitable homography is found, a crop of the augmented image $\text{Crop}(\mathcal{I}^{\text{aug}}_n, \mathbb{P}_c)$ is computed, as well as a crop of the warped augmented image at time $n + t$, $\text{Crop}(\text{Warp}(\mathcal{I}^{\text{aug}}_{n+t}, \mathbf{G}^{-1}), \mathbb{P}_c)$. This keeps all computationally expensive operations outside the loop.

[3] `https://pypi.org/project/Shapely`

Algorithm 2 Homography generation algorithm.

Randomly sample crop polygon $\mathbb{P}_c$ of desired shape in $\mathbb{P}_b$
while rollouts $<$ maximum rollouts **do**
  Randomly sample $\mathbf{G}_{\text{4point}}$, where $\Delta u_i \wedge \Delta v_i \in [-\varrho, \varrho] \, \forall i$
  Perspective transform boundary polygon $p : \mathbb{P}_b, \mathbf{G}^{-1} \rightarrow \mathbb{P}'_b$
  Compute intersection matrix $\mathbf{DE9IM}(\mathbb{P}'_b, \mathbb{P}_c)$
  if DE9IM $= \begin{bmatrix} T & * & * \\ * & * & * \\ F & F & * \end{bmatrix}$ **then**
   return $\mathbf{G}_{\text{4point}}$, $\mathbb{P}_c$
  end if
  rollouts $\leftarrow$ rollouts $+ 1$
end while
  return $\mathbf{0}$, $\mathbb{P}_c$

4.4 Experiments

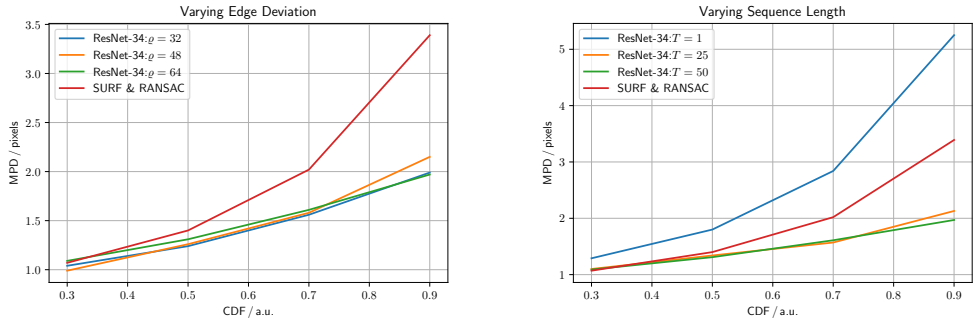
We train DNNs on a 80% train split of the HFR da Vinci® surgery dataset from Fig. 4.1a. The 20% test split is referred to as test set in the following. Inference is performed on the ground truth set from Section 4.3.1.3. We compute the MPD of the predicted value for $\mathbf{G}_{\text{4point}}$ from the desired one. We then compute the CDF of all MPDs. We evaluate the CDF at different thresholds t_i, $i \in \{30, 50, 70.90\}$, e.g. 30% of all homography estimations are below a MPD of t_{30}. We additionally evaluate the compute time on a GeForce RTX 2070 GPU, and a Intel Core i7-9750H central processing unit (CPU).

4.4.1 Backbone Search

In this experiment, we aim to find the best performing backbone for homography estimation. Therefore, we run the same experiment repeatedly with fixed hyper-parameters, and varying backbones. We train each network for 50 epochs, with a batch size of 64, using the Adam optimizer with a learning rate of 2×10^{-4}. The edge devation ϱ is set to 32, and the sequence length T to 25.

4.4.2 Homography Generation Algorithm

In this experiment, we evaluate the *homography generation algorithm*. For this experiment we fix the backbone to a ResNet-34, and train it for 100 epochs, with a batch size of 256, using the Adam optimizer with a learning rate of 1×10^{-3}. Initially, we fix the sequence length T to 25, and train on different edge deviations $\varrho \in \{32, 48, 64\}$. Next, we fix the edge deviation ϱ to 48, and train on different



(a) Varying edge deviation $\varrho \in \{32, 48, 64\}$, and fixed sequence length $T = 25$.

(b) Varying sequence length $T \in \{1, 25, 50\}$, and fixed edge deviation $\varrho = 48$.

Figure 4.4: Homography generation optimization, referring to Section 4.5.2. Shown is a ResNet-34 homography estimation for different homography generation configurations, and a SURF & RANSAC homography estimation for reference. The edge deviation ϱ is varied in (a), and the sequence length T is varied in (b).

sequence lengths $T \in \{1, 25, 50\}$, where a sequence length of 1 corresponds to a static pair of images.

4.5 Results

4.5.1 Backbone Search

The results are listed in Tab. 4.1. It can be seen that the deep methods generally outperform the classical methods on the test set. There is a tendency that models with more parameters perform better. On the ground truth set, this tendency vanishes. The differences in performance become independent of the number of parameters. Noticeably, many backbones still outperform the classical methods across all thresholds on the ground truth set, and low compute regime models also run quicker on CPU than comparable classical methods. E.g. we find that EfficientNet-B0, and RegNetY-400MF run at 36 Hz, and 50 Hz on a CPU, respectively. Both outperform SURF & RANSAC in homography estimation, which runs at 20 Hz.

4.5.2 Homography Generation Algorithm

Given that ResNet-34 performs well on the ground truth set, and executes fast on the GPU, we run the *homography generation algorithm* experiments with it. It can be seen in Fig. 4.4a, that the edge deviation ϱ is neglectable for inference. In Fig. 4.4b, one sees the effects of the sequence length T on the inference performance. Notably, with $T = 1$, corresponding to static image pairs, the SURF & RANSAC

Table 4.1: Results referring to Section 4.5.1. All methods are tested on the da Vinci® HFR test set, indicated by t_i^{test}, and the Cholec80 inference set, indicated by t_i^{gt}. Best, and second best metrics are highlighted with bold character. Improvements in precision $t_{90,\text{imp}}^{\text{gt}}$ and compute time CPU_{imp} are given w.r.t. SURF & RANSAC.

Name	$t_{30}^{\text{test}}/t_{30}^{\text{gt}}$ [pixels]	$t_{50}^{\text{test}}/t_{50}^{\text{gt}}$ [pixels]	$t_{70}^{\text{test}}/t_{70}^{\text{gt}}$ [pixels]	$t_{90}^{\text{test}}/t_{90}^{\text{gt}}$ [pixels]	$t_{90,\text{imp}}^{\text{gt}}$ [%]	params [M]	flops [M]	GPU [ms]	CPU [ms]	CPU_{imp} [%]
VGG-style	4.83/2.45	6.47/2.94	8.68/3.59	13.23/5.41	−60	92.92	11.12	**2** ± 1	83 ± 2	−69 ± 33
ResNet-18	1.42/1.12	1.95/1.33	2.82/1.58	5.06/2.20	35	11.19	6.02	**3** ± 1	31 ± 3	38 ± 13
ResNet-34	1.33/**1.02**	1.81/**1.19**	2.56/**1.52**	4.63/2.08	**39**	21.3	11.74	6 ± 1	51 ± 5	−3 ± 23
ResNet-50	1.40/1.08	1.89/1.33	2.70/1.57	4.79/2.21	35	23.53	13.12	10 ± 1	72 ± 4	−46 ± 29
EfficientNet-B0	1.36/1.09	1.83/1.31	2.62/**1.50**	4.64/**2.01**	**41**	4.02	1.28	12 ± 2	28 ± 2	43 ± 12
EfficientNet-B1	1.32/**1.02**	1.77/**1.26**	2.50/1.57	4.42/**2.01**	**41**	6.52	1.88	17 ± 1	37 ± 1	25 ± 15
EfficientNet-B2	1.40/1.06	1.85/1.29	2.57/1.55	4.42/2.15	37	7.71	2.16	17 ± 2	41 ± 1	18 ± 16
EfficientNet-B3	1.31/1.05	1.75/1.36	2.44/1.68	4.23/2.26	33	10.71	3.14	20 ± 2	55 ± 4	−11 ± 23
EfficientNet-B4	**1.23**/1.08	**1.65**/1.31	**2.29**/1.69	**4.02**/2.14	37	17.56	4.88	24 ± 2	68 ± 5	−38 ± 29
EfficientNet-B5	1.26/1.18	1.67/1.35	**2.30**/1.65	**4.02**/**2.06**	**39**	28.36	7.62	29 ± 2	93 ± 5	−89 ± 37
RegNetY-400MF	1.55/**1.01**	2.07/1.29	2.90/1.60	5.08/2.12	37	3.91	1.32	13 ± 1	**20** ± 1	**58** ± 8
RegNetY-600MF	1.47/1.03	1.98/1.28	2.80/1.56	4.87/2.21	35	5.45	1.94	13 ± 1	24 ± 3	52 ± 11
RegNetY-800MF	1.43/1.08	1.92/1.32	2.70/1.59	4.76/2.12	37	5.50	2.54	12 ± 1	24 ± 1	51 ± 10
RegNetY-1.6GF	1.38/1.03	1.83/1.27	2.52/1.60	4.35/2.16	36	10.32	5.08	21 ± 2	42 ± 4	16 ± 18
RegNetY-4.0GF	1.27/1.05	1.69/**1.26**	2.36/1.66	4.17/2.17	36	19.57	12.36	21 ± 2	66 ± 5	−34 ± 28
RegNetY-6.4GF	**1.21**/1.04	**1.64**/1.27	2.32/1.60	4.17/2.09	38	29.30	19.72	25 ± 3	98 ± 6	−100 ± 40
SURF & RANSAC	4.06/1.07	5.65/1.40	7.93/2.02	13.62/3.39	0	N/A	N/A	N/A	49 ± 9	0 ± 27
SIFT & RANSAC	4.28/1.25	6.02/1.76	8.65/2.48	16.52/4.63	−37	N/A	N/A	N/A	37 ± 9	25 ± 22
ORB & RANSAC	6.52/1.65	10.48/2.47	20.12/3.71	122.66/6.81	−101	N/A	N/A	N/A	**12** ± 2	**76** ± 6

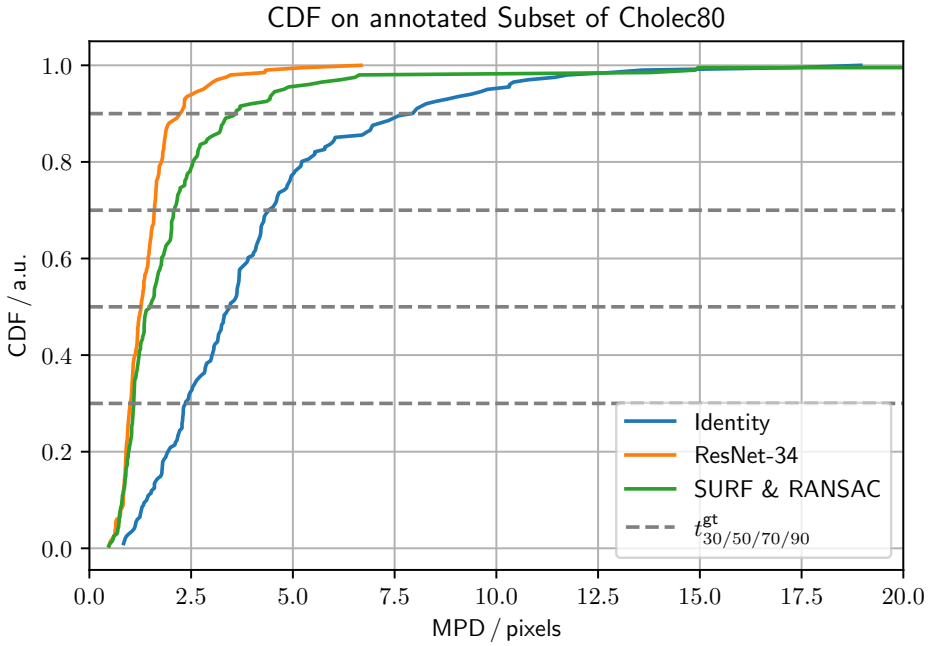


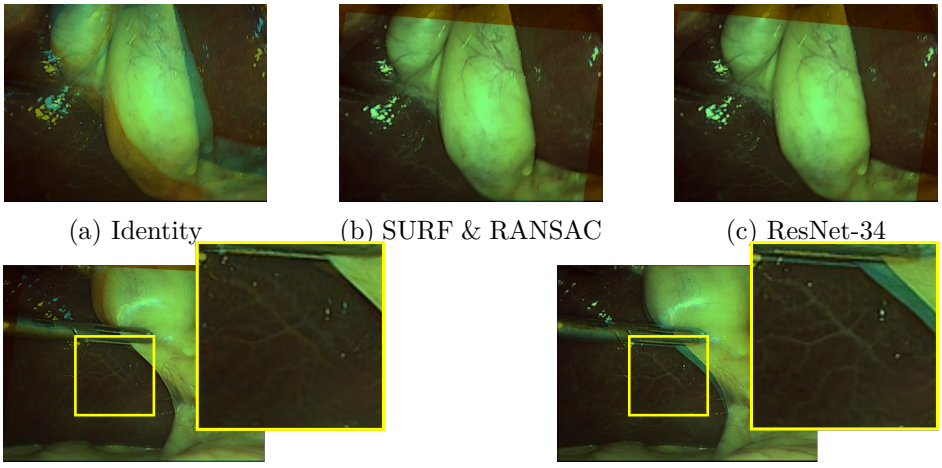
Figure 4.5: CDF for SURF & RANSAC, and ResNet-34, trained with a sequence length $T = 25$, and edge deviation $\varrho = 48$. The identity is added for reference. CDF thresholds for the SURF & RANSAC are $t^{gt}_{1/10/30/50/70/90} = 0.51/0.80/1.09/1.48/2.07/3.53$ pixels, and for the ResNet-34 $t^{gt}_{1/10/30/50/70/90} = 0.50/0.83/1.00/1.26/1.59/2.15$ pixels. ResNet-34 generally performs better, and has no outliers.

homography estimation outperforms the ResNet-34. For the other sequence lengths, ResNet-34 outperforms the classical homography estimation. The CDF for the best performing combination of parameters, with $T = 25$, and $\varrho = 48$, is shown in Fig. 4.5. Our method generally outperforms SURF & RANSAC. The advantage of our method becomes most apparent for a CDF ≥ 0.5. Even the identity outperforms SURF & RANSAC for large MPDs. This aligns with the qualitative observation that motion is often overestimated by SURF & RANSAC, which is shown in Fig. 4.6. An exemplary video is provided[2].

4.6 Conclusion and Future Work

In this work we estimate homographies in dynamic surgical scenes using a supervised learning approach. We train our method on a newly acquired, synthetically

[2] `https://drive.google.com/file/d/1totjHbhIMEL7a-QAiL7B1rT44wvWB6lO/view?usp=sharing`



(a) Identity (b) SURF & RANSAC (c) ResNet-34

(d) Object Motion - SURF & RANSAC, zoom (e) Object Motion - ResNet-34, zoom

Figure 4.6: Classical homography estimation using a SURF feature detector under RANSAC outlier rejection, and the proposed deep homography estimation with a ResNet-34 backbone, referring to Sec. 4.5.2. Shown are blends of consecutive images from a 5 fps resampled Cholec80 exemplary sequence [Twinanda 2017]. Decreasing the framerate from originally 25 fps to 5 fps, increases the motion in between consecutive frames. (Top row) Homography estimation under predominantly camera motion. Both methods perform well. (Bottom row) Homography estimation under predominantly object motion. Especially in the zoomed images it can be seen that the classical method (d) misaligns the stationary parts of the image, whereas the proposed method (e) aligns the background well. This goes to show that the novel data augmentation of Section 4.3.3 principally enables camera motion imitation learning from handheld laparoscopes, see Fig. 1.19.

modified da Vinci® surgery dataset and successfully cross the domain gap to videos of laparoscopic surgeries. To do so, we introduce extensive data augmentation and continuously generate synthetic camera motion through a novel *homography generation algorithm*.

Results Discussion In Section 4.5.1, we find that, despite the domain gap for the ground truth set, DNNs outperform classical methods, which is indicated in Tab. 4.1. The homography estimation performance proves to be independent of the number of model parameters, which indicates an overfit to the test data. The independence of the number of parameters allows to optimize the backbone for computational requirements. E.g., a typical laparoscopic setup runs at $25 - 30$ Hz, the classical method would thus already introduce a bottleneck at 20 Hz. On the other hand, EfficientNet-B0, with 36 Hz, and RegNetY-400MF, with 50 Hz, introduce no latency, and could be integrated into systems without GPU.

In Section 4.5.2, we find that increasing the edge deviation has no effect on the homography estimation, see Fig. 4.4a. This is because the motion in the ground truth set does not exceed the motion in the training set. In Fig. 4.4b, we further find how training DNNs on synthetically modified da Vinci® surgery image sequences enables our method to isolate camera from object and tool motion, validating our method. In Fig. 4.5, it is demonstrated that ResNet-34 generally outperforms SURF & RANSAC. This shows that generating camera motion synthetically through homographies, which approximates the surgical scene as a plane, does not pose an issue.

Future Work The object and tool motion invariant camera motion estimation allows one to extract a laparoscope holder's actions from videos of laparoscopic interventions, which enables the generation of image-action-pairs. In future work, we will generate image-action-pairs from laparoscopic datasets and apply IL to them. Describing camera motion (actions) by means of a homography is grounded in the previous Chapter 3, where we demonstrated that a robotic laparoscope could indeed be controlled by means of a homography under RCM constraint. This work will therefore support the transition towards robotic automation approaches. It might further improve augmented reality, and image mosaicing methods in dynamic surgical environments.

Chapter 5

Self-supervised Laparoscopic Camera Motion Prediction for Imitation Learning

Table of Contents

Disclaimer This Chapter 5 is an *in extenso* reproduction of [Huber 2023b]. Only Section 5.1 was altered to highlight additional context within the scope of this thesis.

5.1 Introduction

In the previous Chapter 4, it was demonstrated that the hypothesized supervised laparoscopic camera motion extraction from Fig. 1.20 indeed outperforms classical means of extraction, see Fig. 4.5 and Fig. 4.6. This was achieved through a novel data augmentation step, refer Fig. 4.3. In fact, the introduced method proved so robust that it outperformed classical methods on MIS data, although it was trained on RMIS data, see Fig. 4.6. This is necessary for extracting homographies from MIS data, i.e. non-robotic data, and to imitate a handheld laparoscope's motion. The extracted homography, i.e. action $\hat{a}_t^*$, refer Fig. 1.19, could already be combined with the proposed visual servo scheme from Chapter 3, however, this would ignore that the surgical scene changes over time (temporality assumption) as was mentioned in Section 3.1.

In this chapter Chapter 5, we will finally investigate how the extracted homographies can be utilized in a self-supervised training scheme to implement the IL for predicting desired actions $\hat{a}_t^*$ from image states $\hat{s}_t$, see Fig. 1.19. This builds on the realization that camera motion is intrinsic to videos of laparoscopic interventions and that camera motion can be harvested from them, as was already discussed in Section 1.6.2.2. Since no public dataset exists for state-action pairs, some work continues to focus on the tools to infer camera motion [Li 2021a], or learns on a robotic setup altogether [Li 2022b] where camera motion is accessible. In this chapter, we address those shortcomings, to enable IL on large scale datasets of retrospective laparoscopic surgery videos.

5.1.1 Contributions

We build on Chapter 4 for computationally efficient state-action pair extraction from publicly available datasets of laparoscopic interventions, which yields more than 20× the amount of data that was used in the closed source data of [Li 2022a]. To our best knowledge, this dataset marks the largest and only state-action pair dataset for laparoscopic camera motion IL. On top of the sheer amount of additional data, we extract homography estimates at runtime, de facto generating data on the fly. Contrary to [Li 2022a], our camera motion extraction does not rely on image features, which are sparse in surgical videos, and is intrinsically capable to differentiate between camera and object motion. We further propose a novel importance sampling and data augmentation step for achieving camera motion automation IL in this chapter. It is indeed the importance sampling, which renders learning to predict motion feasible.

5.2 Materials and Methods

The proposed approach to learning camera motion prediction is summarized in Fig. 5.1. The following sections will describe its key components in more detail.

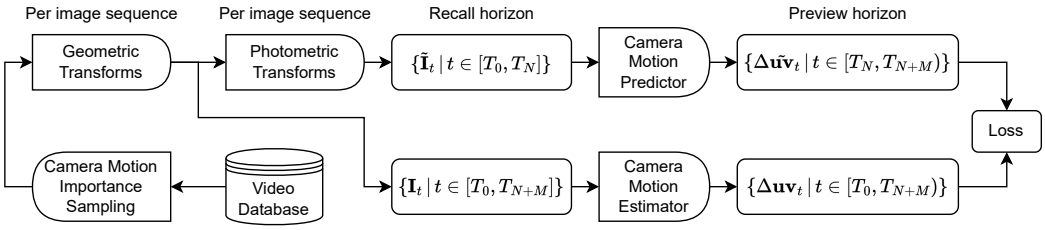


Figure 5.1: Training pipeline, refer to Section 5.2.3. From left to right: Image sequences are importance sampled from the video database and random augmentations are applied per sequence online. The lower branch estimates camera motion between subsequent frames, which is taken as pseudo-ground-truth for the upper branch, which learns to predict camera motion on a preview horizon.

5.2.1 Theoretical Background

Points on a plane, as observed from a moving camera, transform by means of the 3×3 projective homography matrix $\mathbf{G}$ in image space, see Section 1.6.2.3. Thus, predicting future camera motion (up to scale) may be equivalently treated as predicting future projective homographies.

It has been shown in [DeTone 2016] that the four point representation of the projective homography, *i.e.,* taking the difference between four points in homogeneous coordinates $\Delta\mathbf{uv} = \{\mathbf{p}_i - \mathbf{p}'_i \mid i \in [0, 4)\} \in \mathbb{R}^{4\times 2}$ that are related by $\mathbf{G}\mathbf{p}_i \sim \mathbf{p}'_i \; \forall i$, is better suited for deep learning applications than the 3×3 matrix representation of a homography, which is harder to estimate correctly. Therefore, in this work, we treat camera motion $\mathcal{C}$ as a sequence of four point homographies on a time horizon $[T_0, T_{N+M})$, N being the recall horizon's length, M being the preview horizon's length. Time points lie Δt apart, that is $T_{i+1} = T_i + \Delta t$. For image sequences of length N+M, we work with four point homography sequences $\mathcal{C} = \{\Delta\mathbf{uv}_t \mid t \in [T_0, T_{N+M})\}$.

5.2.2 Data and Data Preparation

Three datasets are curated to train and evaluate the proposed method: two cholecystectomy datasets (laparoscopic gallbladder removal), namely Cholec80 [Twinanda 2017] and HeiChole [Wagner 2023], and one hysterectomy dataset (laparoscopic uterus removal), namely AutoLaparo [Wang 2022].

To remove status indicator overlays from the laparoscopic videos, which may hinder the camera motion estimator, we identify the bounding circle of the circular field of view using [Budd 2022]. We crop the view about the center point of the bounding circle to a shape of 240×320, so that no black regions are prominent in the images.

All three datasets are split into training, validation, and testing datasets. We split the videos by frame count into $80 \pm 1\,\%$ training and $20 \pm 1\,\%$ testing. Training and testing videos never intersect. We repeat this step to further split the training dataset into (pure) training and validation datasets.

Due to errors during processing the raw data, i.e. failure to detect the bounding circle automatically, we exclude videos 19, 21, and 23 from HeiChole, as well as videos 22, 40, 65, and 80 from Cholec80. This results in dataset sizes of: Cholec80 - $4.4e6$ frames at 25 fps, HeiChole - $9.5e5$ frames at 25 fps, and AutoLaparo - $7.1e4$ frames at 25 fps.

5.2.3 Proposed Pipeline

5.2.3.1 Video Database and Importance Sampling

The curated data from Section 5.2.2 is accumulated into a video database. Image sequences of length $N + M$ are sampled at a frame increment of Δn between subsequent frames and with Δc frames between the sequence's initial frames. Prior to adding the videos to the database, an initial offline run is performed to estimate camera motion $\mathbf{\Delta uv}$ between the frames. This creates image-motion correspondences of the form $(\mathbf{I}_n\,, \mathbf{I}_{n+\Delta n}\,, \mathbf{\Delta uv}_n)$. Image-motion correspondences where $\mathbb{E}(||\mathbf{\Delta uv}_n||_2) > \sigma$, with sigma being the standard deviation over all motions in the respective dataset, define anchor indices n. Image sequences are sampled such that the last image in the recall horizon lies at index $n = N - 1$, marking the start of a motion. The importance sampling samples indices from the intersection of all anchor indices, shifted by $-N$, with all possible starting indices for image sequences.

5.2.3.2 Geometric and Photometric Transforms

The importance sampled image sequences are fed to a data augmentation stage. This stage entails geometric and photometric transforms. The distinction is made because downstream, the pipeline is split into two branches. The upper branch serves as camera motion prediction whereas the lower branch serves as camera motion estimation, also refer to the next section. As it acts as the source of pseudo-ground-truth, it is crucial that the camera motion estimator performs under optimal conditions, hence no photometric transforms, i.e. transforms that change brightness

/ contrast / fog etc., are applied. Photometrically transformed images shall further be denoted as $\tilde{\mathbf{I}}$. To encourage same behavior under different perspectives, geometric transforms are applied, i.e. transforms that change orientation / up to down / left to right etc. Transforms are always sampled randomly, and applied consistently to the entire image sequence.

5.2.3.3 Camera Motion Estimator and Predictor

The goal of this work is to have a predictor learn camera motion computed by an estimator. The predictor takes as input a photometrically and geometrically transformed recall horizon $\{\tilde{\mathbf{I}}_t \,|\, t \in [T_0, T_N)\}$ of length N, and predicts camera motion $\tilde{\mathcal{C}} = \{\Delta\tilde{\mathbf{uv}}_t \,|\, t \in [T_N, T_{N+M})\}$ on the preview horizon of length M. The estimator takes as input the geometrically transformed preview horizon $\{\mathbf{I}_t \,|\, t \in [T_M, T_{N+M})\}$ and estimates camera motion $\mathcal{C}$, which serves as a target to the predictor. The estimator is part of the pipeline to facilitate on-the-fly perspective augmentation via the geometric transforms.

5.3 Experiments and Evaluation Methodology

The following two sections elaborate the experiments we conduct to investigate the proposed pipeline from Fig. 5.1 in Section 5.2.3. First the camera motion estimator is investigated, followed by the camera motion predictor.

5.3.1 Camera Motion Estimator

5.3.1.1 Camera Motion Distribution

To extract the camera motion distribution, we run the camera motion estimator of Chapter 4 with a ResNet-34 backbone over all datasets from Section 5.2.2. We map the estimated four point homographies to up/down/left/right/zoom-in/zoom-out for interpretability. Left/right/up/down corresponds to all four point displacements $\Delta\mathbf{uv}$ consistently pointing left/right/up/down respectively. Zoom-in/out corresponds to all four point displacements $\Delta\mathbf{uv}$ consistently pointing inwards/outwards. Rotation left corresponds to all four point displacements pointing up right, bottom right, and so on. Same for rotation right. Camera motion is defined static if it lies below the standard deviation in the dataset. The frame increment is set to 0.25 s, corresponding to $\Delta n = 5$ for the 25 fps videos.

5.3.1.2 Online Camera Motion Estimation

Since the camera motion estimator is executed online, memory footprint and computational efficiency are of importance. Therefore, we evaluate the estimator of Chapter 4 with a ResNet-34 backbone, SURF & RANSAC, and local feature matching (LoFTR) [Sun 2021] & RANSAC. Each estimator is run 1000 times on a single image sequence of length $N + M = 15$ with an NVIDIA GeForce RTX 2070 GPU and an Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz.

5.3.2 Camera Motion Predictor

5.3.2.1 Model Architecture

For all experiments, the camera motion predictor is a ResNet-18/34/50, with the number of input features equal to the recall horizon N $\times$ 3 (RGB), where N = 14. We set the preview horizon M = 1. The frame increment is set to 0.25 s, or $\Delta n = 5$ for the 25 fps videos. The number of frames between clips is also set to 0.25 s, or $\Delta c = 5$.

5.3.2.2 Training Details

The camera motion predictor is trained on each dataset from Section 5.2.2 individually. For training on Cholec80/HeiChole/AutoLaparo, we run 80/50/50 epochs on a batch size of 64 with a learning rate of $2.5e-5/1.e-4/1.e-4$. The learning rates for Cholec80 and HeiChole relate approximately to the dataset's training sizes, see 5.2. For Cholec80, we reduce the learning rate by a factor 0.5 at epochs 50, 75. For Heichole/AutoLaparo we drop the learning rate by a factor 0.5 at epoch 35. The loss in Fig. 5.1 is set to the mean pairwise distance between estimation and prediction $\mathbb{E}(||\Delta\tilde{\mathbf{uv}}_t - \Delta\mathbf{uv}_t||_2) + \lambda\mathbb{E}(||\Delta\tilde{\mathbf{uv}}_t||_2)$ with a regularizer that discourages the identity $\Delta\tilde{\mathbf{uv}}_t = \mathbf{0}$ (i.e. no motion). We set $\lambda = 0.1$.

5.3.2.3 Evaluation Metrics

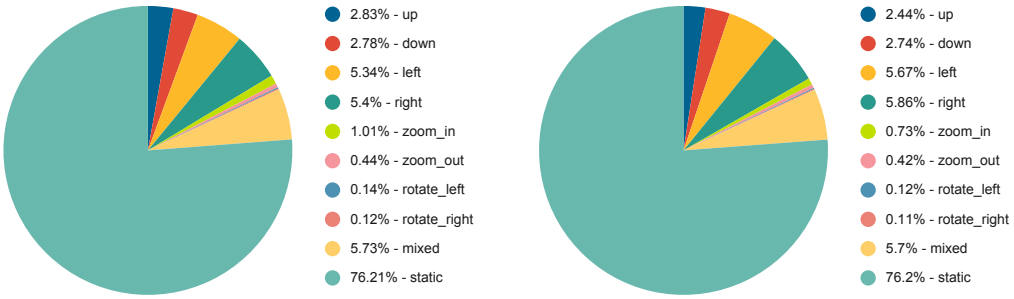
For evaluation we compute the mean pairwise distance between estimated and predicted motion $\mathbb{E}(||\Delta\tilde{\mathbf{uv}}_t - \Delta\mathbf{uv}_t||_2)$. All camera motion predictors are benchmarked against a baseline, that is a $\mathcal{O}(1)/\mathcal{O}(2)$-Taylor expansion of the estimated camera motion $\Delta\mathbf{uv}_t$. Furthermore, the model that is found to perform best is evaluated on the multi-class labels (left, right, up, down) that are provided in AutoLaparo.

5.4 Results

5.4.1 Camera Motion Estimator

5.4.1.1 Camera Motion Distribution

The camera motion distributions for all datasets are shown in Fig. 5.2. It is observed that for a large fraction of the sequences there is no significant camera motion (Cholec80 76.21%, HeiChole 76.2%, AutoLaparo 71.29%). This finding supports the importance sampling that was introduced in Section 5.2.3.1. It can further be seen that e.g. left/right and up/down motions are equally distributed.



(a) Cholec80 (b) HeiChole

Figure 5.2: Camera motion distribution, refer to Section 5.3.1. AutoLaparo: 2.81% - up, 1.88% - down, 4.48% - left, 3.38% - right, 0.45% - zoom_in, 0.2% - zoom_out, 0.3% - rotate_left 0.3%, - rotate_right 14.9% - mixed, 71.29% - static.

5.4.1.2 Online Camera Motion Estimation

The results of the online camera motion estimation are summarized in Table 5.1. The deep homography estimation with a Resnet34 backbone executes 11× quicker and has the lowest GPU memory footprint of the GPU accelerated methods. This allows for efficient implementation of the proposed online camera motion estimation in Fig. 5.1.

Table 5.1: Memory footprint and execution time of different camera motion estimators, refer to Section 5.3.1.2.

Method	Execution time [s]	Speed-up [a.u.]	Model / Batch [Mb]
Resnet34	$\mathbf{0.016 \pm 0.048}$	**11.1**	**664/457**
LoFTR & RANSAC	0.178 ± 0.06	1.0	669/2412
SURF & RANSAC	0.131 ± 0.024	1.4	NA

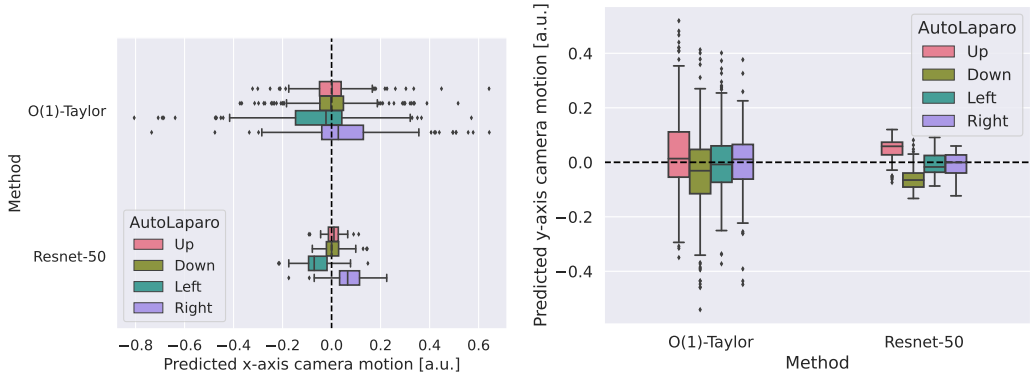
5.4.2 Camera Motion Prediction

The camera motion prediction results for all datasets are highlighted in Table 5.2. It can be seen that significant improvements over the baseline are achieved on the Cholec80 and HeiChole datasets. Whilst the learned prediction performs better on average than the baseline, no significant improvement is found for the AutoLaparo dataset.

The displacement of the image center point under the predicted camera motion for AutoLaparo is plotted against the provided multi-class motion annotations and shown in Fig. 5.3. It can be seen that the camera motion predictions align well with the ground truth labels.

Table 5.2: Camera motion predictor performance, refer to Section 5.3.2. Taylor baselines predict based on previous estimated motion, ResNets based on images.

Dataset	Train Size [Frames]	Mean Pairwise Distance [Pixels]				
		Taylor		ResNet (proposed)		
		$\mathcal{O}(1)$	$\mathcal{O}(2)$	18	34	50
Cholec80	3.5e6	27.2 ± 23.1	36.4 ± 31.2	$\mathbf{14.8} \pm 11.7$	$\mathbf{14.4} \pm 11.4$	$\mathbf{14.4} \pm 11.4$
HeiChole	7.6e5	29.7 ± 26.4	39.8 ± 35.9	$\mathbf{15.8} \pm 12.5$	$\mathbf{15.8} \pm 12.5$	$\mathbf{15.8} \pm 12.5$
AutoLaparo	5.9e4	19.4 ± 18.4	25.8 ± 24.7	$\mathbf{11.2} \pm 11.0$	$\mathbf{11.3} \pm 11.0$	$\mathbf{11.3} \pm 11.0$



(a) Predicted camera motion along x-axis, scaled by image size to $[-1, 1]$.

(b) Predicted camera motion along y-axis, scaled by image size to $[-1, 1]$.

Figure 5.3: Predicted camera motion on AutoLaparo, refer to Section 5.3.2. Camera motion predictor trained on Cholec80 with ResNet-50 backbone, see Table 5.2. Shown is the motion of the image center under the predicted homography. Clearly, for videos labeled left/right, the center point is predicted to move left/right and for up/down labels, the predicted left/right motion is centered around zero (a). Same is observed for up/down motion in (b), where left/right motion is zero-centered.

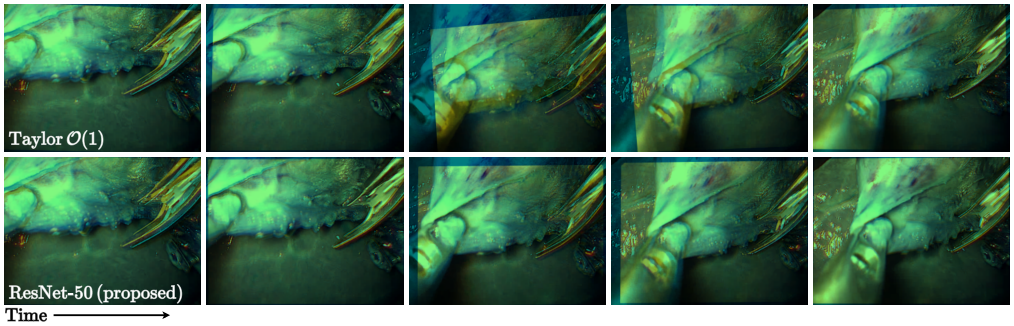


Figure 5.4: Exemplary camera motion prediction, refer to Section 5.3.2. In the image sequence, the attention changes from the right to the left tool. We warp the past view (yellow) by the predicted homography and overlay the current view (blue). Good alignment corresponds to good camera motion prediction. Contrary to the baseline, the proposed method predicts the motion well. Data taken from HeiChole test set, ResNet-50 backbone trained on Cholec80, refer Table 5.2.

5.5 Conclusion and Future Work

Results Discussion To the best of our knowledge, this work is the first to demonstrate that camera motion can indeed be learned from retrospective videos of laparoscopic interventions, with no manual annotation. Self-supervision is achieved by harvesting image-motion correspondences using a camera motion estimator, see Fig. 5.1. The camera motion predictor is shown to generate statistically significant better predictions over a baseline in Table 5.2 as measured using pseudo-ground-truth and on multi-class manually annotated motion labels from AutoLaparo in Fig. 5.3. An exemplary image sequence in Fig. 5.4 demonstrates successful camera motion prediction on HeiChole. These results were achieved through the key finding from Fig. 5.2, which states that most image sequences, i.e. static ones, are irrelevant to learning camera motion. Consequentially, we contribute a novel importance sampling method, as described in Section 5.2.3.1. Finally, we hope that our open-source commitment will help the community explore this area of research further.

Limitations A current limitations of this work is the preview horizon M of length 1. One might want to extend it for model predictive control. Furthermore, to improve explainability to the surgeon, but also to improve the prediction in general, it would be beneficial to include auxiliary tasks, e.g. tool and organ segmentation, surgical phase recognition, and audio. There also exist limitations for the camera motion estimator. The utilized camera motion estimator is efficient and isolates object motion well from camera motion, but is limited to relatively small camera

motions. Improving the camera motion estimator to large camera motions would help increase the preview horizon M.

Future Work In future work, we will execute this model in a real setup for investigating transferability. This endeavor is backed by Chapter 3, which demonstrates how the learned homography could immediately be deployed on a robotic laparoscope holder. It might prove necessary to fine-tune the presented policy through reinforcement learning from human feedback (RLHF).

Chapter 6

Conclusion and Future Work

Table of Contents

In this chapter, we highlight some of the main contributions that were achieved within this thesis. We further discuss the major shortcomings as well as future prospectives that could address those limitations.

6.1 Summary

In the introductory Chapter 1, we analyzed the inevitable rise of robot assisted laparoscopy and linked it to surgeon benefits and future prospects, Section 1.3.1. We identified spatial awareness and automation as key targets for enhancing and alleviating, driving factors and roadblocks, respectively, see Fig. 1.10. To address these, we proposed marker-free unified calibration for enhanced spatial awareness with improved clinical workflow in Section 1.4.3.2. Moreover, in Fig. 1.19, we outlined a framework for learning to imitate camera motion from a camera-assistant-held laparoscope, see Fig. 1.4. We argued that, through a mixture of classical control, and IL in image space, it might be possible to imitate a surgeon through a robot laparoscope holder despite their different embodiment. The following sections will discuss the extend to which the proposed solutions met the pinpointed targets, and further suggest future research directions when targets where not fully met or could be improved upon.

6.2 Marker-free Unified Eye-hand Calibration

Contributions In Chapter 2, we introduce a novel marker-free unified eye-hand calibration procedure. The method solely relies on stereo RGB-D images, robot mesh or CAD files, and joint position readings, see Fig. 2.3. We contribute an algorithm for robust point-to-plane ICP in Algorithm 1. Extensive comparisons of single-shot capabilities against classical eye-in- and eye-to-hand as well as hand-shake calibrations are demonstrated on system, see Fig. 2.4, and on-par results are summarized in Table 2.2 with visually good alignments in Fig. 2.6. We finally deploy an improved version of the method in a clinically relevant scenario, Fig. 2.5, and showcase qualitatively accurate renders in Fig. 2.8, with close to perfect IoUs in Table 2.1, and corresponding minor segmentation-render differences in Fig. 2.9.

Shortcomings and Future Work The proposed robust ICP already exhibits convincing registration results and impressive scalability to any industrial serial manipulator with, compared to the differentiable rendering alternatives, little dependencies. In the clinical context, however, draping poses a procedural necessity that renders the proposed approach infeasible. As such, future work should focus

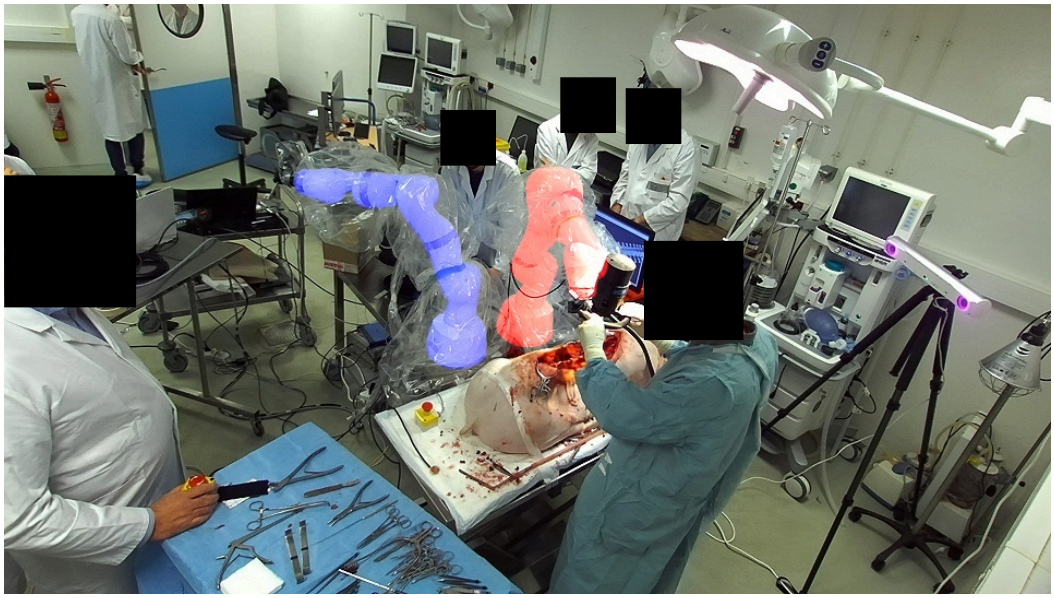


Figure 6.1: Render of robots, given the registration results of Section 2.6.2, overlayed on view in draped stage. Note that the camera drifted slightly and the registration was not corrected for in the above, yet. Hyperspectral camera robot (left, blue) and drilling robot (right, red), see Fig. 2.5. Refers to Section 6.2.

on marker-free registration despite draping. To this end, we contribute a dataset of draped robots with ground-truth registrations, see Fig. 6.1. The collected dataset could be used to train a segmentation model on ignoring the draping, thus facilitating registration via methods, such as differentiable rendering. The learned segmentation model could further be used to alleviate the need for manual segmentation, see Fig. 2.1, that is currently required to prompt SAM, which poses a significant challenge for clinical translation.

Key Takeaways In Chapter 2, we address the need for improved spatial awareness with unaltered clinical workflows, Section 1.3.3.1, Fig. 1.10. We contribute a novel dataset for first steps towards clinical translation on draped systems. Moving forward, this research should enhance modular systems Section 1.3.2, Table 1.1, with accompanying benefits such as reduced cost and operation time. Serial manipulators in industrial applications could benefit through the presented approach immediately.

6.3 Homography-based Visual Servo with Remote Center of Motion Constraint

Contributions In Chapter 3, we introduced a novel image-based visual servoing approach with RCM constraint. We attempted to shift the control paradigm from a tool-centric control policy towards a view-centric control policy to address the flaws of the dominant visual servos, refer Section 1.5.2, Fig. 1.13. To this end, we derived a homography-based visual servo which does not rely on depth nor on tool distance for inferring actions Fig. 3.1. We did so, introducing a projection operator for mapping target camera frame velocities to available DoF, (3.14). We then deployed the newly derived control in a novel view-graph-based semi-autonomous scheme on a real system, see Fig. 3.2, Fig. 1.18. The proposed method demonstrated good image space convergence properties throughout traversing the view graph from current to target view whilst retained a deviation from the target RCM below 5 mm, see Fig. 3.3. The proposed method further indicated robustness against patient repositioning, i.e. coordinate system change, see Table 3.1.

Shortcomings and Future Work The most apparent shortcoming of the proposed method is the assumption of a mostly static surgical scene which clearly does not hold, see Fig. 3.4. This temporality assumption, however, is not so crucial within the overarching context of the thesis, as ultimately the executed action is not graph-based but rather incrementally sampled from a learned expert policy $\pi_{\mathrm{E}} : \hat{s}_t \rightarrow \hat{a}_t^*$, refer Section 1.5.5. A shortcoming that weighs much heavier is e.g. the lack of force sensing. No trocar-laparoscope external forces nor arm-environment external forces are incorporated into the controller. Arm-environment external forces could e.g. be minimized through nullspace projection and the redundant DoF of the manipulator. The proposed controller furthermore does not explicitly constrain solutions to the RCM, and convergence to undesired minima might occur. Further research should thus introduce constrained optimization instead.

Key Takeaways Altogether, it was demonstrated that indeed the introduced visual servo could serve as means for executing actions $\hat{a}_t^*$ for the hypothesized pipeline, refer Fig. 1.19, but that future improvements would be necessary for successful clinical translation.

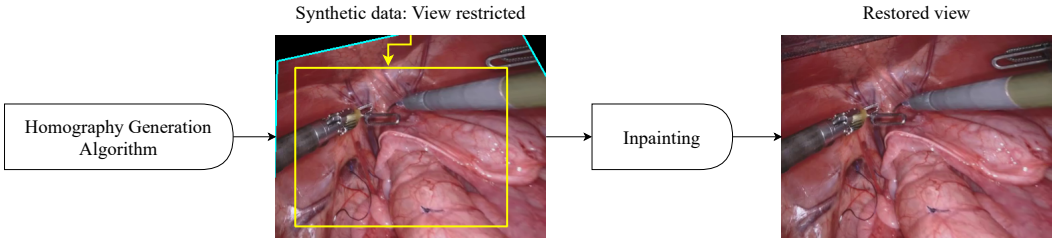


Figure 6.2: Proposed inpainting for data retrieval. The *homography generation algorithm* from Section 4.3.3, Fig. 4.3, introduces black boundaries and thus restricted views. Generative inpainting could help restore the entire view. Fourier inpainting done via [Suvorov 2021], **not** fine-tuned. Refers to Section 6.4.

6.4 Homography-based Camera Motion Estimation

Contributions In Chapter 4, we introduce a novel data augmentation Algorithm 2 for runtime generation of synthethic camera motion, see Fig. 4.3. We introduce a new dataset of camera motion separated da Vinci® surgery image sequences through exploitation of the clamping mechanism, which we initially proposed in Fig. 1.20. Therefore, we utilize all available HFR datasets from Table 1.3 and summarize their respective sizes in Fig. 4.1. We conduct an extensive backbone search for models that are regressed onto the synthetic data across multiple compute regimes and devices, Table 4.1. For evaluating the transferability from videos of RMIS to MIS, we hand-annotate static landmarks in video sequences of laparoscopic interventions Fig. 4.2. We find that models which are supervisedly trained on the novel dataset with the novel *homography generation algorithm* outperform classical feature-based homography estimators. We find that they only outperform classical methods when trained on sequence lengths $N > 1$, see Fig. 4.4, justifying our proposed approach. We further find that the learned models outperform the feature-based ones across varying edge deviations ϱ. A quantitative statistical significance is shown in Fig. 4.5. Qualitatively, we find that static vessels in the background are aligned better through our novel approach, Fig. 4.6.

Shortcomings and Future Work Whilst significant improvements are demonstrated over SOTA homography estimates, there still exist shortcomings of this work. The *homography generation algorithm*, although capable of generating indefinite synthetic camera motion, introduces the necessity to crop the view as otherwise black borders would be introduced. This intrinsically limits the magnitude of synthetic camera motion and therefore also the range of learnable camera motion. This fact is somewhat underevaluated through the hand-annotations from Fig. 4.2

as images were annotated on a frame-to-frame basis with relatively little camera motion in-between. For future work, we suggest to train a generative inpainting model that restores the entire view post camera motion generation, see Fig. 6.2. In this example, an unrefined inpainting model was utilized, showcasing the feasibility of dedicated methods. We suggest the use of a standard generative approach as opposed to diffusion models, such as [Rombach 2022], since the iterative process might not satisfy the runtime requirements of the *homography generation algorithm*. Furthermore, future research might aim at incorporating depth information for camera motion synthesis, e.g. through [Budd 2024], which is available on GitHub[1].

Key Takeaways Chapter 4 introduced several novelties for estimating camera motion in dynamic surgical scenes despite the presence of tool and organ motion, proving the hypothesized pipeline in Fig. 1.20. As such, the work presented in Chapter 4 presents the SOTA methodology for generation of state-action pairs $(\hat{s}_t, \hat{a}_t^*)$, refer Section 1.5.5, to be used in the IL proposal of Fig. 1.19.

6.5 Homography-based Camera Motion Prediction

Contributions In Chapter 5, we introduce a novel approach for self-supervised camera motion IL from retrospective videos of laparoscopic surgeries, refer Fig. 5.1. The proposed method utilizes the camera motion extraction of Chapter 4 to generate state-action pairs at runtime, for which efficiency is demonstrated in Table 5.1. This approach allows us to introduce the separation of photometric and geometric transforms, such that behaviors can be learned regardless of the relative patient positioning, as we can still extract camera motion on geometrically transformed sequences. The proposed method leverages a novel importance sampling, which is grounded on an analysis of camera motion distribution over large-scale video datasets of cholecystectomies, see Fig. 5.2. Without this importance sampling, only identity policies would be learned. We then contribute the first large-scale IL on publicly available data and find significantly improved performance over baselines Table 5.2. We reaffirm this finding through predicting camera motion on the AutoLaparo dataset [Wang 2022], and showcase that the predicted motion aligns with the provided labels Fig. 5.3. A qualitative prediction on a video sequence is shown in Fig. 5.4.

Shortcomings and Future Work While this work demonstrates that camera motion can be predicted in a fully self-supervised manner, it does still have short-

[1] github.com/charliebudd/transferring-relative-monocular-depth-to-surgical-vision

comings. One of the major shortcomings that were not addressed is the limited preview horizon $M = 1$, see Section 5.3.2.1. As such, the model only predicts 0.25 s in advance and could, e.g. not be used for model predictive control. It further only has a context window, i.e. the recall horizon, of $N = 14$ or 3.5 s. For increased context window sizes, it might be crucial that the camera motion estimator, Fig. 5.1, is capable of estimating larger motions, which is an inherited issue of Section 6.4, and could be addressed through Fig. 6.2. The model is thus mostly suited for predicting how motion will continue but currently lacks bootstrapping capabilities. Part of the reason for lacking bootstrapping capabilities can further be found in the partial state observance, see Fig. 1.19. The proposed method only accesses iamges $\hat{s}_t$, and does not have any prior procedural knowledge, which we argued was beyond scope of this thesis, nor knows about device readings or surroundings in general, which are inaccessible in the public datasets. However, with the advent of massively pre-trained large language models, such as Llama [Touvron 2023], Mistral [Jiang 2023], and GPT-4 [OpenAI 2024], the incorporation of procedureal knowledge might become feasible, as can already be observed in tasks such as visual question answering [Seenivasan 2022]. Ultimately, this might lead to foundation-like models for laparoscopic camera motion action prediction. For these types of models, although we discarded simulation-based RL in Section 1.5.4, RLHF might become a relevant research direction, which is much more sample efficient than pure RL. Admittance controllers with RCM, which are suitable for RLHF, can e.g. be found in Appendix A.

Key Takeaways In Chapter 5, we found first indicators that actions $\hat{a}_t^*$ might indeed be an immediately learnable task, similar to the auxiliary tasks of Fig. 1.15. We thus closed the loop for the hypothesized pipeline of Fig. 1.19, but argued that further research would be necessary.

6.6 Closing Remarks

In this thesis, entitled *Data-driven Robotic Endoscope Automation*, we embarked on a journey to address level five autonomy as introduced in the Foreword (Section 1.1). Through a thorough exploration of various domains—clinical, economic, and technical—we formulated a holistic approach to embodiment-invariant IfO, illustrated in Fig. 1.19 (Section 1.6.2). Implementing this conceptual framework led to the development of novel solutions discussed in detail across sections from Section 6.2 to Section 6.5. It's important to note that while level five autonomy remains elusive, significant strides have been made, some of which could alleviate surgical

staff of mundane tasks. As thus, it was shown how marker-free eye-hand calibration could enable spatially aware autonomous robotic laparoscope holders in Chapter 2, allowing for the safe execution of autonous controls, as well as other prospects such as workspace optimization and quicker setup times in general. The following Chapter 3 then explored visual servoing towards target views via homography-based image registration under RCM constraint, already allowing surgeons to control robotic system in a semi-autonomous and collaborative fashion. Finally, Chapter 4 successfully generated state-action pairs for learning to predict camera motion in image space in Chapter 5, although this work demonstrated much more difficult than anticipated, and only initial progress was made. What is currently missing is the deployment of the learned policy to a phantom setup, but achieving this proved beyond the scope of this thesis, since IL turned out a hard problem on data of handheld laparoscopes, and should be explored in the future.

This work not only sheds light on promising avenues for future research but also underscores the emergence of new challenges and opportunities.

In the spirit of fostering a beginner's mindset, we conclude with a quote from the book Zen Mind, Beginner's Mind:

In the beginner's mind there are many possibilities, but in the expert's there are few.

— Shunryu Suzuki [Suzuki 1970]

As such, we envision this thesis serving as a catalyst, fostering innovation and inspiring fresh perspectives in approaching this complex problem of laparoscopic camera motion automation.

Appendix A

LBR-Stack: ROS 2 and Python Integration of KUKA FRI for Med and IIWA Robots

---

Table of Contents

Disclaimer This Appendix A is an *in extenso* reproduction of [Huber 2023a].

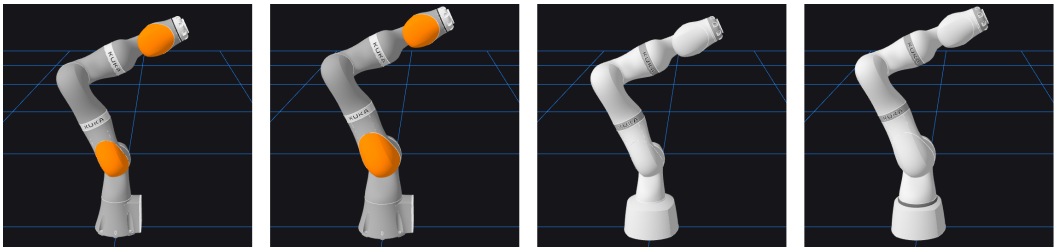


Figure A.1: Supported robots in the LBR-Stack. From left to right: KUKA LBR IIWA7, IIWA14, Med7, Med14. Visualizations made using Foxglove [1].

A.1 Summary

The `LBR-Stack`[2] is a collection of packages that simplify the usage and extend the capabilities of KUKA's Fast Robot Interface (FRI) [Schreiber 2010]. It is designed for mission critical hard real-time applications. Supported are the `KUKA LBR Med7/14` and `KUKA LBR IIWA7/14` robots in the Gazebo simulation [Koenig 2004] and for communication with real hardware. A demo video can be found here. An overview of the software architecture is shown in Figure A.2.

At the `LBR-Stack`'s core are two packages:

- **fri**: Integration of KUKA's original FRI client library into CMake.
- **fri_vendor**: Vendor library that integrates the **fri** into the ROS 2 build sytem.

All other packages are built on top. These include Python bindings and packages for integration into the ROS and ROS 2:

- **pyFRI**: Python bindings for the **fri**.
- **lbr_fri_ros2_stack**: ROS 1/2 integration of the `KUKA LBR`s through the **fri_vendor**.

For brevity, and due to the architectural advantages over ROS [Macenski 2022], only ROS 2 is considered in the following. The **lbr_fri_ros2_stack** comprises the following packages:

- **lbr_bringup**: Python library for launching the different components.
- **lbr_description**: Description files for the `Med7/14` and `IIWA7/14` robots.
- **lbr_demos**: Demonstrations for simulation and the real robots.

[1] Foxglove: `https://foxglove.dev/ros`.
[2] LBR-Stack: `https://github.com/lbr-stack/`

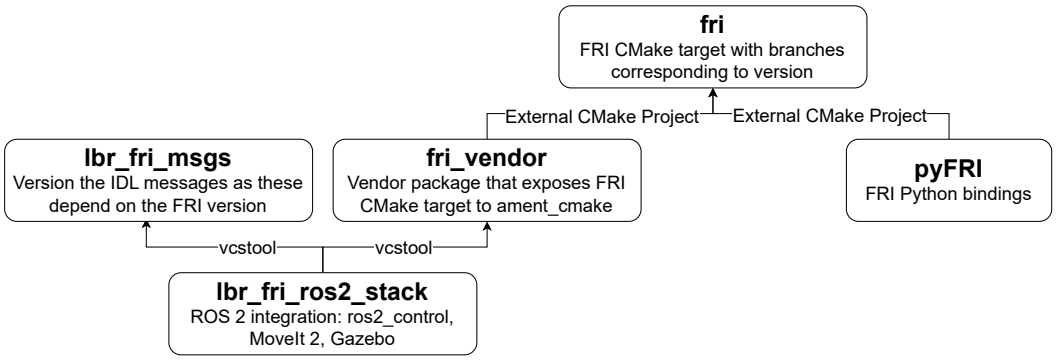


Figure A.2: An overview of the overall software architecture. There exists a single source for KUKA's FRI. This design facilitates that downstream packages, i.e. the Python bindings and the ROS 2 package, can easily support multiple FRI versions. The ROS 2 side utilizes vcstool[3].

- **lbr_fri_msgs**: Interface Definition Language (IDL) equivalent of FRI protocol buffers.
- **lbr_fri_ros2**: FRI ROS 2 interface through `realtime_tools` [Chitta 2017].
- **lbr_ros2_control**: Interface and controllers for `ros2_control` [Magyar 2023].
- **lbr_moveit_config**: MoveIt 2 configurations [Coleman 2014].

A.2 Statement of need

An overview of existing work that interfaces the KUKA LBRs from an external computer is given in Table 1. We broadly classify these works into custom communication solutions [Hennersperger 2016; Safeea 2018; Serrano-Muñoz 2023] and communication solutions through KUKA's FRI UDP channel [Bednarczyk 2023; Chatzilygeroudis 2019]. The former can offer greater flexibility while the latter offer a well defined interface and direct software support from KUKA. Contrary to the custom communication solutions, the FRI solutions additionally enable hard real-time communication, that is beneficial for mission critical development. Stemming from translational medical research, this work therefore focuses on the FRI.

Limitations with the current FRI solutions are:

1. Only support `IIWA7/14` robots, not `Med7/14`.
2. Do not provide Python bindings.

[3]vcstool: `https://github.com/dirk-thomas/vcstool`.

3. Maintainability:
    - Modified client source code iiwa_ros.
    - FRI client library tangled into source code iiwa_ros2.
4. Partial support of FRI functionality. Both, iiwa_ros and iiwa_ros2, exclusively aim at providing implementations of the ROS 1/2 hardware abstraction layer. This does not support:
    - FRI's cartesian impedance control mode.
    - FRI's cartesian control mode (FRI version 2 and above).

The first original contribution of this work is to add support for the `KUKA LBR Med7/14` robots, which, to the best author's knowledge, does not exist in any other work. The second novel contribution of this work is to provide Python bindings. This work solves the maintainability by outsourcing the FRI into the separate **fri** and **fri_vendor** packages, which leaves the FRI's source code untouched and simply provides build support. 4. is solved by defining an IDL message to KUKA's `nanopb` command and state protocol buffers in **lbr_fri_msgs**. These messages can then be interfaced from ROS 1/2 topics or from the ROS 1/2 hardware abstraction layer.

Table A.1: Overview of existing frameworks for interfacing the KUKA LBRs. A bullet point indicates support for the respective feature.

Framework	IIWA	Med	ROS	ROS 2	Hard Real-time	FRI	pyFRI	Position Control	Impedance Control	Cartesian Impedance Control	Hardware Interface
lbr-stack	•	•	•	•		•	•	•	•	•	•
iiwa_ros	•		•		•	•		•	•		•
iiwa_ros2	•			•	•	•		•	•		•
iiwa-stack	•		•					•	•	•	
libiiwa	•		•	•				•	•	•	
KST-KUKA	•							•	•	•	

A.3 Acknowledgement

We want to acknowledge the work in [Hennersperger 2016], as their MoveIt configurations were utilized in a first iteration of this project.

Bibliography

[Abdelaal 2020] Alaa Eldin Abdelaal, Nancy Hong, Apeksha Avinash, Divya Budihal, Maram Sakr, Gregory D. Hager, and Septimiu E. Salcudean. *Orientation Matters: 6-DoF Autonomous Camera Movement for Minimally Invasive Surgery*. 2020. arXiv: 2012.02836 [cs.RO] (cit. on p. 59).

[Aghakhani 2013] Nastaran Aghakhani, Milad Geravand, Navid Shahriari, Marilena Vendittelli, and Giuseppe Oriolo. "Task control with remote center of motion constraint for minimally invasive robotic surgery". In: *2013 IEEE International Conference on Robotics and Automation*. IEEE. 2013, pp. 5807–5812. DOI: 10.1109/ICRA.2013.6631412 (cit. on pp. 60, 102–104).

[Agrawal 2018] Ankur Agrawal. "Automating endoscopic camera motion for teleoperated minimally invasive surgery using inverse reinforcement learning". PhD thesis. Master's thesis, Worcester Polytechnic Institute, 2018 (cit. on p. 63).

[Agustinos 2014] A. Agustinos, R. Wolf, J. A. Long, P. Cinquin, and S. Voros. "Visual servoing of a robotic endoscope holder based on surgical instrument tracking". In: *5th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics*. 2014, pp. 13–18. DOI: 10.1109/BIOROB.2014.6913744 (cit. on p. 59).

[Ahmidi 2017] Narges Ahmidi, Lingling Tao, Shahin Sefati, Yixin Gao, Colin Lea, Benjamín Béjar Haro, Luca Zappella, Sanjeev Khudanpur, René Vidal, and Gregory D. Hager. "A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic

Surgery". In: *IEEE Transactions on Biomedical Engineering* 64.9 (2017), pp. 2025–2041. DOI: 10.1109/TBME.2016.2647680 (cit. on pp. 70, 76).

[Alhusseinawi 2023] Hayder Alhusseinawi, Rikke Haase, Sten Rasmussen, Jørgen B. Jensen, and Pernille S. Kingo. "Validation of a surgical workspace scale during robot-assisted surgery". In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 19.1 (2023), e2482. DOI: 10.1002/rcs.2482. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/rcs.2482 (cit. on p. 49).

[Allan 2019] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, Luis Herrera, Wenqi Li, Vladimir Iglovikov, Huoling Luo, Jian Yang, Danail Stoyanov, Lena Maier-Hein, Stefanie Speidel, and Mahdi Azizian. *2017 Robotic Instrument Segmentation Challenge*. 2019. arXiv: 1902.06426 [cs.CV] (cit. on pp. 61, 76, 120).

[Allan 2020] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, Avinash Kori, Varghese Alex, Ganapathy Krishnamurthi, David Rauber, Robert Mendel, Christoph Palm, Sophia Bano, Guinther Saibro, Chi-Sheng Shih, Hsun-An Chiang, Juntang Zhuang, Junlin Yang, Vladimir Iglovikov, Anton Dobrenkii, Madhu Reddiboina, Anubhav Reddy, Xingtong Liu, Cong Gao, Mathias Unberath, Myeonghyeon Kim, Chanho Kim, Chaewon Kim, Hyejin Kim, Gyeongmin Lee, Ihsan Ullah, Miguel Luna, Sang Hyun Park, Mahdi Azizian, Danail Stoyanov, Lena Maier-Hein, and Stefanie Speidel. *2018 Robotic Scene Segmentation Challenge*. 2020. arXiv: 2001.11190 [cs.CV] (cit. on pp. 76, 120).

[Allard 2007] Jérémie Allard, Stéphane Cotin, François Faure, Pierre-Jean Bensoussan, François Poyer, Christian

Duriez, Hervé Delingette, and Laurent Grisoni. "Sofa-an open source framework for medical simulation". In: *MMVR 15-Medicine Meets Virtual Reality*. Vol. 125. IOP Press. 2007, pp. 13–18 (cit. on p. 63).

[Archana 2018] Arumugom Archana, Sathasivam Sureshkumar, Chellappa Vijayakumar, Chinnakali Palanivel, and Palanivel Chinnakali. "Comparing the harmonic scalpel with electrocautery in reducing postoperative flap necrosis and seroma formation after modified radical mastectomy in carcinoma breast patients: a double-blind prospective randomized control trail". In: *Cureus* 10.4 (2018). PMID: 29904617. DOI: 10.7759/cureus.2476 (cit. on p. 40).

[Attanasio 2021] Aleks Attanasio, Bruno Scaglioni, Elena De Momi, Paolo Fiorini, and Pietro Valdastri. "Autonomy in Surgical Robotics". In: *Annual Review of Control, Robotics, and Autonomous Systems* 4.Volume 4, 2021 (2021), pp. 651–679. ISSN: 2573-5144. DOI: 10.1146/annurev-control-062420-090543. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-control-062420-090543 (cit. on p. 36).

[Attia 2017] Mohamed Attia, Mohammed Hossny, Saeid Nahavandi, and Hamed Asadi. "Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder". In: *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2017, pp. 3373–3378. DOI: 10.1109/SMC.2017.8123151 (cit. on p. 61).

[Aytar 2018] Yusuf Aytar, Tobias Pfaff, David Budden, Thomas Paine, Ziyu Wang, and Nando de Freitas. "Playing hard exploration games by watching YouTube". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/

2018/file/35309226eb45ec366ca86a4329a2b7c3-Paper.pdf (cit. on p. 66).

[Azizian 2014] Mahdi Azizian, Mahta Khoshnam, Nima Najmaei, and Rajni V Patel. "Visual servoing in medical robotics: a survey. Part I: endoscopic and direct vision imaging–techniques and applications". In: *The international journal of medical robotics and computer assisted surgery* 10.3 (2014). PMID: 24106103, pp. 263–274. DOI: 10.1002/rcs.1531 (cit. on p. 58).

[Bakalar 2021] Nicholas Bakalar. *Are Robotic Surgeries Really Better?* https://www.nytimes.com/2021/08/16/well/live/robotic-surgery-benefits.html. Accessed: 2024-02-14. 2021 (cit. on p. 44).

[Bakari 2007] Mohamed J. Bakari, Khaled M. Zied, and Derek W. Seward. "Development of a Multi-Arm Mobile Robot for Nuclear Decommissioning Tasks". In: *International Journal of Advanced Robotic Systems* 4.4 (2007), p. 51. DOI: 10.5772/5665 (cit. on p. 78).

[Bano 2020] Sophia Bano, Francisco Vasconcelos, Marcel Tella-Amo, George Dwyer, Caspar Gruijthuijsen, Emmanuel Vander Poorten, Tom Vercauteren, Sebastien Ourselin, Jan Deprest, and Danail Stoyanov. "Deep learning-based fetoscopic mosaicking for field-of-view expansion". In: *International journal of computer assisted radiology and surgery* 15.11 (2020), pp. 1807–1816. DOI: 10.1007/s11548-020-02242-8 (cit. on p. 119).

[Bardozzo 2022] Francesco Bardozzo, Toby Collins, Antonello Forgione, Alexandre Hostettler, and Roberto Tagliaferri. "StaSiS-Net: A stacked and siamese disparity estimation network for depth reconstruction in modern 3D laparoscopy". In: *Medical Image Analysis* 77 (2022), p. 102380. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102380. URL: https://www.sciencedirect.com/science/article/pii/S1361841522000329 (cit. on p. 62).

[Battaglia 2021] Edoardo Battaglia, Jacob Boehm, Yi Zheng, Andrew R. Jamieson, Jeffrey Gahan, and Ann Majewicz Fey. "Rethinking Autonomous Surgery: Focusing on Enhancement over Autonomy". In: *European Urology Focus* 7.4 (2021), pp. 696–705. ISSN: 2405-4569. DOI: 10.1016/j.euf.2021.06.009. URL: https://www.sciencedirect.com/science/article/pii/S2405456921001711 (cit. on p. 36).

[Bawa 2020] Vivek Singh Bawa, Gurkirt Singh, Francis KapingA, Inna Skarga-Bandurova, Alice Leporini, Carmela Landolfo, Armando Stabile, Francesco Setti, Riccardo Muradore, Elettra Oleari, and Fabio Cuzzolin. *ESAD: Endoscopic Surgeon Action Detection Dataset*. 2020. arXiv: 2006.07164 [cs.CV] (cit. on pp. 76, 120).

[Bay 2006] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. "SURF: Speeded Up Robust Features". In: *Computer Vision – ECCV 2006*. Ed. by Aleš Leonardis, Horst Bischof, and Axel Pinz. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 404–417. ISBN: 978-3-540-33833-8. DOI: 10.1007/11744023_32 (cit. on p. 108).

[Bednarczyk 2023] M. Bednarczyk and J. H. G. Guzmán. *ROS 2 stack for KUKA iiwa collaborative robots*. https://github.com/ICube-Robotics/iiwa_ros2. 2023 (cit. on p. 151).

[Benhimane 2006] S. Benhimane and E. Malis. "Homography-based 2D visual servoing". In: *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* 2006, pp. 2397–2402. DOI: 10.1109/ROBOT.2006.1642061 (cit. on pp. 73, 105, 106).

[Blausencom 2014] Blausen.com. "Medical gallery of Blausen Medical". In: *WikiJournal of Medicine* 1 (2014). ISSN: 2002-4436. DOI: DOI:10.15347/wjm/2014.010 (cit. on p. 39).

[Bodenstedt 2018] Sebastian Bodenstedt, Max Allan, Anthony Agustinos, Xiaofei Du, Luis Garcia-Peraza-Herrera, Hannes

Kenngott, Thomas Kurmann, Beat Müller-Stich, Sebastien Ourselin, Daniil Pakhomov, Raphael Sznitman, Marvin Teichmann, Martin Thoma, Tom Vercauteren, Sandrine Voros, Martin Wagner, Pamela Wochner, Lena Maier-Hein, Danail Stoyanov, and Stefanie Speidel. *Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery*. 2018. arXiv: 1805.02475 [cs.CV] (cit. on p. 120).

[Bodenstedt 2019a] Sebastian Bodenstedt, Dominik Rivoir, Alexander Jenke, Martin Wagner, Michael Breucha, Beat Müller-Stich, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. "Active learning using deep Bayesian networks for surgical workflow analysis". In: *International journal of computer assisted radiology and surgery* 14.6 (2019), pp. 1079–1087. DOI: 10.1007/s11548-019-01963-9 (cit. on p. 62).

[Bodenstedt 2019b] Sebastian Bodenstedt, Martin Wagner, Lars Mündermann, Hannes Kenngott, Beat Müller-Stich, Michael Breucha, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. "Prediction of laparoscopic procedure duration using unlabeled, multimodal sensor data". In: *International journal of computer assisted radiology and surgery* 14.6 (2019), pp. 1089–1095. DOI: 10.1007/s11548-019-01966-6 (cit. on p. 62).

[Brantner 2021] Gerald Brantner and Oussama Khatib. "Controlling Ocean One: Human–robot collaboration for deep-sea manipulation". In: *Journal of Field Robotics* 38.1 (2021), pp. 28–51. DOI: 10.1002/rob.21960. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21960 (cit. on p. 78).

[Budd 2022] Charlie Budd, Luis C. Garcia-Peraza Herrera, Martin Huber, Sebastien Ourselin, and Tom Vercauteren. "Rapid and robust endoscopic content area estimation: a lean GPU-based pipeline and curated benchmark dataset". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and*

Visualization (2022). ISSN: 21681171. DOI: 10.1080/21681163.2022.2156393 (cit. on pp. 121, 134).

[Budd 2023a] Charlie Budd, Luis C. Garcia-Peraza Herrera, Martin Huber, Sebastien Ourselin, and Tom Vercauteren. "Rapid and robust endoscopic content area estimation: a lean GPU-based pipeline and curated benchmark dataset". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 11.4 (2023), pp. 1215–1224. DOI: 10.1080/21681163.2022.2156393 (cit. on p. 18).

[Budd 2023b] Charlie Budd, Jianrong Qiu, Oscar MacCormac, Martin Huber, Christopher Mower, Mirek Janatka, Théo Trotouin, Jonathan Shapey, Mads S. Bergholt, and Tom Vercauteren. "Deep Reinforcement Learning Based System for Intraoperative Hyperspectral Video Autofocusing". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Cham: Springer Nature Switzerland, 2023, pp. 658–667. ISBN: 978-3-031-43996-4 (cit. on p. 18).

[Budd 2024] Charlie Budd and Tom Vercauteren. *Transferring Relative Monocular Depth to Surgical Vision with Temporal Consistency*. 2024. arXiv: 2403.06683 [cs.CV] (cit. on pp. 62, 146).

[Caron 2018] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. "Deep Clustering for Unsupervised Learning of Visual Features". In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Cham: Springer International Publishing, 2018, pp. 139–156. ISBN: 978-3-030-01264-9. DOI: 10.1007/978-3-030-01264-9_9 (cit. on p. 66).

[Cartucho 2021] João Cartucho, Samyakh Tukra, Yunpeng Li, Daniel S. Elson, and Stamatia Giannarou. "VisionBlender: a tool to efficiently generate computer vision datasets

for robotic surgery". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 9.4 (2021), pp. 331–338. DOI: 10.1080/21681163.2020.1835546 (cit. on p. 64).

[Chatterjee 2024] Swastika Chatterjee, Soumyajit Das, Karabi Ganguly, and Dibyendu Mandal. "Advancements in robotic surgery: innovations, challenges and future prospects". In: *Journal of Robotic Surgery* 18.1 (2024), p. 28. DOI: 10.1007/s11701-023-01801-w (cit. on p. 50).

[Chatzilygeroudis 2019] K. Chatzilygeroudis, M. Mayr, B. Fichera, and A. Billard. *Iiwa_ros: A ROS stack for KUKA's IIWA robots using the fast research interface*. http://github.com/epfl-lasa/iiwa_ros. 2019 (cit. on p. 151).

[Chen 2023] Linghao Chen, Yuzhe Qin, Xiaowei Zhou, and Hao Su. "EasyHeC: Accurate and Automatic Hand-Eye Calibration Via Differentiable Rendering and Space Exploration". In: *IEEE Robotics and Automation Letters* 8.11 (Nov. 2023), pp. 7234–7241. ISSN: 2377-3774. DOI: 10.1109/lra.2023.3315551 (cit. on pp. 56, 80).

[Chitta 2017] Sachin Chitta, Eitan Marder-Eppstein, Wim Meeussen, Vijay Pradeep, Adolfo Rodríguez Tsouroukdissian, Jonathan Bohren, David Coleman, Bence Magyar, Gennaro Raiola, Mathias Lüdtke, and Enrique Fernandez Perdomo. "ros_control: A generic and simple control framework for ROS". In: *Journal of Open Source Software* 2.20 (2017), p. 456. DOI: 10.21105/joss.00456 (cit. on p. 151).

[Cimen 2019] Haci Ibrahim Cimen, Yavuz Tarik Atik, Serkan Altinova, Oztug Adsan, and Mevlana Derya Balbay. "Does the experience of the bedside assistant effect the results of robotic surgeons in the learning curve of robot assisted radical prostatectomy?" In: *International braz j urol* 45 (2019), pp. 54–60. DOI: 10.1590/S1677-5538.IBJU.2018.0184 (cit. on p. 49).

[Clementini 1994] Eliseo Clementini, Jayant Sharma, and Max J. Egenhofer. "Modelling topological spatial relations: Strategies for query processing". In: *Computers & Graphics* 18.6 (1994), pp. 815–822. ISSN: 0097-8493. DOI: 10.1016/0097-8493(94)90007-8. URL: https://www.sciencedirect.com/science/article/pii/0097849394900078 (cit. on p. 123).

[Col 2020] Tommaso Da Col, Andrea Mariani, Anton Deguet, Arianna Menciassi, Peter Kazanzides, and Elena De Momi. "SCAN: System for Camera Autonomous Navigation in Robotic-Assisted Surgery". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2020, pp. 2996–3002. DOI: 10.1109/IROS45743.2020.9341548 (cit. on pp. 58, 59).

[Coleman 2014] David Coleman, Ioan Sucan, Sachin Chitta, and Nikolaus Correll. *Reducing the Barrier to Entry of Complex Robotic Software: a MoveIt! Case Study*. 2014. arXiv: 1404.3785 [cs.RO] (cit. on p. 151).

[Costa Rocha 2019] Cristian da Costa Rocha, Nicolas Padoy, and Benoit Rosa. "Self-supervised surgical tool segmentation using kinematic information". In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 8720–8726. DOI: 10.1109/ICRA.2019.8794334 (cit. on p. 61).

[Csirzó 2023] Ádám Csirzó, Dénes Péter Kovács, Anett Szabó, Péter Fehérvári, Árpád Jankó, Péter Hegyi, Péter Nyirády, Zoltán Sipos, Levente Sára, Nándor Ács, et al. "Robot-assisted laparoscopy does not have demonstrable advantages over conventional laparoscopy in endometriosis surgery: a systematic review and meta-analysis". In: *Surgical Endoscopy* (2023), pp. 1–11. DOI: 10.1007/s00464-023-10587-9 (cit. on p. 44).

[Cuschieri 1994] A Cuschieri, S Shimi, S Banting, LK Nathanson, and A Pietrabissa. "Intraoperative cholangiography during laparoscopic cholecystectomy: routine vs selective policy". In: *Surgical endoscopy* 8 (1994),

pp. 302–305. DOI: 10.1089/lps.1993.3.27 (cit. on p. 40).

[Czempiel 2020] Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. "TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks". In: *Lecture Notes in Computer Science*. Springer International Publishing, 2020, pp. 343–352. ISBN: 9783030597160. DOI: 10.1007/978-3-030-59716-0_33 (cit. on p. 62).

[Davenport 2019] Thomas Davenport and Ravi Kalakota. "The potential for artificial intelligence in healthcare". In: *Future Hospital Journal* 6 (June 2019), pp. 94–98. DOI: 10.7861/futurehosp.6-2-94 (cit. on p. 50).

[Dergachyova 2016] Olga Dergachyova, David Bouget, Arnaud Huaulmé, Xavier Morandi, and Pierre Jannin. "Automatic data-driven real-time segmentation and recognition of surgical workflow". In: *International Journal of Computer Assisted Radiology and Surgery* 11.6 (2016), pp. 1081–1089. DOI: 10.1007/s11548-016-1371-x (cit. on p. 62).

[DeTone 2016] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. *Deep Image Homography Estimation*. 2016. arXiv: 1606.03798 [cs.CV] (cit. on pp. 119, 122, 133).

[Do 2012] Hyun Min Do, Chanhun Park, and Jin Ho Kyung. "Dual arm robot for packaging and assembling of IT products". In: *2012 IEEE International Conference on Automation Science and Engineering (CASE)* (2012), pp. 1067–1070. DOI: 10.1109/CoASE.2012.6386417 (cit. on p. 78).

[Ellis 2016] R. Darin Ellis, Anthony J. Munaco, Luke A. Reisner, Michael D. Klein, Anthony M. Composto, Abhilash K. Pandya, and Brady W. King. "Task analysis of laparoscopic camera control schemes". In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 12.4 (2016), pp. 576–584. DOI: 10.1002/rcs.1716. URL: https://

onlinelibrary.wiley.com/doi/abs/10.1002/rcs.1716 (cit. on p. 60).

[Eslamian 2016] Shahab Eslamian, L. Reisner, B. King, and A. Pandya. "Towards the Implementation of an Autonomous Camera Algorithm on the da Vinci Platform". In: *Studies in health technology and informatics* 220 (2016). PMID: 27046563, pp. 118–23. DOI: 10.13140/RG.2.2.25637.91364 (cit. on p. 59).

[Eslamian 2017] Shahab Eslamian, Luke A Reisner, Brady W King, and Abhilash K Pandya. *An autonomous camera system using the da vinci research kit*. 2017 (cit. on p. 59).

[Eslamian 2020] Shahab Eslamian, Luke A. Reisner, and Abhilash K. Pandya. "Development and evaluation of an autonomous camera control algorithm on the da Vinci Surgical System". In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 16.2 (2020), e2036. DOI: 10.1002/rcs.2036. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rcs.2036 (cit. on p. 59).

[Esteva 2019] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. "A guide to deep learning in healthcare". In: *Nature medicine* 25.1 (2019), pp. 24–29. DOI: 10.1038/s41591-018-0316-z (cit. on p. 69).

[Farber 2021] S Harrison Farber, Mark A Pacult, Jakub Godzik, Corey T Walker, Jay D Turner, Randall W Porter, and Juan S Uribe. "Robotics in spine surgery: a technical overview and review of key concepts". In: *Frontiers in Surgery* 8 (2021), p. 578674. DOI: 10.3389/fsurg.2021.578674 (cit. on p. 51).

[Finn 2016] Chelsea Finn, Ian Goodfellow, and Sergey Levine. *Unsupervised Learning for Physical Interaction through Video Prediction*. 2016. arXiv: 1605.07157 [cs.LG] (cit. on p. 65).

[Finn 2017a] Chelsea Finn and Sergey Levine. "Deep visual foresight for planning robot motion". In: *2017 IEEE*

International Conference on Robotics and Automation (ICRA). 2017, pp. 2786–2793. DOI: 10.1109/ICRA.2017.7989324 (cit. on p. 65).

[Finn 2017b] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. *One-Shot Visual Imitation Learning via Meta-Learning*. 2017. arXiv: 1709.04905 [cs.LG] (cit. on p. 65).

[Fiorini 2022] Paolo Fiorini, Ken Y Goldberg, Yunhui Liu, and Russell H Taylor. "Concepts and Trends in Autonomy for Robot-Assisted Surgery". In: *Proceedings of the IEEE* 110.7 (2022), pp. 993–1011. DOI: 10.1109/JPROC.2022.3176828 (cit. on p. 49).

[Fosch-Villaronga 2021] Eduard Fosch-Villaronga, Pranav Khanna, Hadassah Drukarch, and Bart HM Custers. "A human in the loop in surgery automation". In: *Nature Machine Intelligence* 3.5 (2021), pp. 368–369. DOI: 10.1038/s42256-021-00349-4 (cit. on p. 36).

[Fuji Tsang 2022] Clement Fuji Tsang, Maria Shugrina, Jean Francois Lafleche, Towaki Takikawa, Jiehan Wang, Charles Loop, Wenzheng Chen, Krishna Murthy Jatavallabhula, Edward Smith, Artem Rozantsev, Or Perel, Tianchang Shen, Jun Gao, Sanja Fidler, Gavriel State, Jason Gorski, Tommy Xiang, Jianing Li, Michael Li, and Rev Lebaredian. *Kaolin: A Pytorch Library for Accelerating 3D Deep Learning Research*. https://github.com/NVIDIAGameWorks/kaolin. 2022 (cit. on p. 80).

[Funke 2018] Isabel Funke, Alexander Jenke, Sören Torge Mees, Jürgen Weitz, Stefanie Speidel, and Sebastian Bodenstedt. "Temporal Coherence-based Self-supervised Learning for Laparoscopic Workflow Analysis". In: *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Cham: Springer International Publishing, 2018, pp. 85–93. ISBN: 978-3-030-01201-4. DOI: 10.1007/978-3-030-01201-4_11 (cit. on p. 62).

[García-Peraza-Herrera 2017a] Luis C. García-Peraza-Herrera, Wenqi Li, Lucas Fidon, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, and Sébastien Ourselin. "ToolNet: Holistically-nested real-time segmentation of robotic surgical tools". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2017, pp. 5717–5722. DOI: 10.1109/IROS.2017.8206462 (cit. on p. 61).

[García-Peraza-Herrera 2017b] Luis C. García-Peraza-Herrera, Wenqi Li, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, and Sébastien Ourselin. "Real-Time Segmentation of Non-rigid Surgical Tools Based on Deep Learning and Tracking". In: *Computer-Assisted and Robotic Endoscopy*. Cham: Springer International Publishing, 2017, pp. 84–95. ISBN: 978-3-319-54057-3. DOI: 10.1007/978-3-319-54057-3_8 (cit. on p. 61).

[Garcia-Peraza-Herrera 2021] Luis C Garcia-Peraza-Herrera, Lucas Fidon, Claudia D'Ettorre, Danail Stoyanov, Tom Vercauteren, and Sébastien Ourselin. "Image Compositing for Segmentation of Surgical Tools Without Manual Annotations". In: *IEEE Transactions on Medical Imaging* 40.5 (2021). PMID: 33556005, pp. 1450–1460. DOI: 10.1109/TMI.2021.3057884 (cit. on p. 61).

[Giannarou 2013] Stamatia Giannarou, Marco Visentini-Scarzanella, and Guang-Zhong Yang. "Probabilistic Tracking of Affine-Invariant Anisotropic Regions". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 130–143. DOI: 10.1109/TPAMI.2012.81 (cit. on pp. 76, 120).

[Glira 2015] Philipp Glira, Norbert Pfeifer, Christian Briese, and Camillo Ressl. "A Correspondence Framework for ALS Strip Adjustments based on Variants of the ICP Algorithm". In: *Photogrammetrie - Fernerkundung - Geoinformation* 2015.4 (Aug. 2015),

pp. 275–289. DOI: 10.1127/pfg/2015/0270 (cit. on p. 85).

[Gomes 2019] Sara Gomes, Maria Teresa Valério, Marta Salgado, Hélder P. Oliveira, and António Cunha. "Unsupervised Neural Network for Homography Estimation in Capsule Endoscopy Frames". In: *Procedia Computer Science* 164 (2019). CENTERIS 2019 - International Conference on ENTERprise Information Systems / ProjMAN 2019 - International Conference on Project MANagement / HCist 2019 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2019, pp. 602–609. ISSN: 1877-0509. DOI: 10.1016/j.procs.2019.12.226. URL: https://www.sciencedirect.com/science/article/pii/S1877050919322732 (cit. on p. 119).

[Green 1984] P. J. Green. "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.2 (1984), pp. 149–170. DOI: 10.1111/j.2517-6161.1984.tb01288.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1984.tb01288.x (cit. on p. 90).

[Gruijthuijsen 2022] Caspar Gruijthuijsen, Luis C. Garcia-Peraza-Herrera, Gianni Borghesan, Dominiek Reynaerts, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Emmanuel Vander Poorten. "Robotic Endoscope Control Via Autonomous Instrument Tracking". In: *Frontiers in Robotics and AI* 9 (Apr. 2022). ISSN: 2296-9144. DOI: 10.3389/frobt.2022.832208 (cit. on pp. 59, 61).

[Gupta 2023] Vishal Gupta. "How to achieve the critical view of safety for safe laparoscopic cholecystectomy: Technical aspects". In: *Annals of Hepato-biliary-pancreatic*

Surgery 27.2 (2023), pp. 201–210. DOI: 10.14701/ahbps.22-064 (cit. on p. 41).

[Hameed 2016] Saad Hameed and Osman Hasan. "Towards autonomous collision avoidance in surgical robots using image segmentation and genetic algorithms". In: *2016 IEEE Region 10 Symposium (TENSYMP)*. 2016, pp. 266–270. DOI: 10.1109/TENCONSpring.2016.7519416 (cit. on p. 49).

[Hasan 2021] Md. Kamrul Hasan, Lilian Calvet, Navid Rabbani, and Adrien Bartoli. "Detection, segmentation, and 3D pose estimation of surgical tools using convolutional neural networks and algebraic geometry". In: *Medical Image Analysis* 70 (2021), p. 101994. ISSN: 1361-8415. DOI: 10.1016/j.media.2021.101994. URL: https://www.sciencedirect.com/science/article/pii/S1361841521000402 (cit. on p. 62).

[Hattab 2020] Georges Hattab, Marvin Arnold, Leon Strenger, Max Allan, Darja Arsentjeva, Oliver Gold, Tobias Simpfendörfer, Lena Maier-Hein, and Stefanie Speidel. "Kidney edge detection in laparoscopic image data for computer-assisted surgery". In: *International journal of computer assisted radiology and surgery* 15.3 (2020), pp. 379–387. DOI: 10.1007/s11548-019-02102-0 (cit. on p. 120).

[Hausman 2017] Karol Hausman, Yevgen Chebotar, Stefan Schaal, Gaurav Sukhatme, and Joseph Lim. *Multi-Modal Imitation Learning from Unstructured Demonstrations using Generative Adversarial Nets*. 2017. arXiv: 1705.10479 [cs.RO] (cit. on p. 65).

[Hennersperger 2016] Christoph Hennersperger, Bernhard Fuerst, Salvatore Virga, Oliver Zettinig, Benjamin Frisch, Thomas Neff, and Nassir Navab. "Towards MRI-based autonomous robotic US acquisitions: a first feasibility study". In: *IEEE transactions on medical imaging* 36.2 (2016), pp. 538–548. DOI: 10.1109/TMI.2016.2620723 (cit. on pp. 151, 154).

[Ho 2016] Jonathan Ho and Stefano Ermon. *Generative Adversarial Imitation Learning*. 2016. arXiv: 1606.03476 [cs.LG] (cit. on p. 65).

[Hoeller 2023] David Hoeller, Nikita Rudin, Dhionis Sako, and Marco Hutter. *ANYmal Parkour: Learning Agile Navigation for Quadrupedal Robots*. 2023. arXiv: 2306.14874 [cs.RO] (cit. on p. 63).

[Hongwei Li 2016] Lingtao Yu an Hongwei Li, Lingyan Zhao, Sixu Ren, and Qing Gu. "Automatic guidance of laparoscope based on the region of interest for robot assisted laparoscopic surgery". In: *Computer Assisted Surgery* 21.sup1 (2016), pp. 17–21. DOI: 10.1080/24699322.2016.1240309 (cit. on p. 59).

[Horaud 1995] Radu Horaud and Fadi Dornaika. "Hand-Eye Calibration". In: *The International Journal of Robotics Research* 14.3 (1995), pp. 195–210. DOI: 10.1177/027836499501400301 (cit. on pp. 55, 80).

[Horváth 2017] Gergely Horváth and Gábor Erdős. "Point cloud based robot cell calibration". In: *CIRP Annals* 66.1 (2017), pp. 145–148. ISSN: 0007-8506. DOI: 10.1016/j.cirp.2017.04.044. URL: https://www.sciencedirect.com/science/article/pii/S0007850617300446 (cit. on p. 80).

[Huang 2021] Baoru Huang, Jian-Qing Zheng, Anh Nguyen, David Tuch, Kunal Vyas, Stamatia Giannarou, and Daniel S. Elson. "Self-supervised Generative Adversarial Network for Depth Estimation in Laparoscopic Images". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Cham: Springer International Publishing, 2021, pp. 227–237. ISBN: 978-3-030-87202-1. DOI: 10.1007/978-3-030-87202-1_22 (cit. on p. 62).

[Huang 2022] Baoru Huang, Anh Nguyen, Siyao Wang, Ziyang Wang, Erik Mayer, David Tuch, Kunal Vyas, Stamatia Giannarou, and Daniel S. Elson. "Simultaneous Depth Estimation and Surgical Tool Segmentation in Laparoscopic Images". In: *IEEE Transactions on Medical Robotics and Bionics* 4.2 (2022),

pp. 335–338. DOI: 10.1109/TMRB.2022.3170215 (cit. on p. 62).

[Huber 2021] Martin Huber, John Bason Mitchell, Ross Henry, Sébastien Ourselin, Tom Vercauteren, and Christos Bergeles. "Homography-based visual servoing with remote center of motion for semi-autonomous robotic endoscope manipulation". In: *2021 International Symposium on Medical Robotics (ISMR)*. IEEE. 2021, pp. 1–7. DOI: 10.1109/ISMR48346.2021.9661563 (cit. on pp. 17, 101).

[Huber 2022] Martin Huber, Sébastien Ourselin, Christos Bergeles, and Tom Vercauteren. "Deep homography estimation in dynamic surgical scenes for laparoscopic camera motion extraction". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 10.3 (2022), pp. 321–329. DOI: 10.1080/21681163.2021.2002195 (cit. on pp. 17, 117).

[Huber 2023a] Martin Huber, Christopher E. Mower, Sebastien Ourselin, Tom Vercauteren, and Christos Bergeles. *LBR-Stack: ROS 2 and Python Integration of KUKA FRI for Med and IIWA Robots*. 2023. arXiv: 2311.12709 [cs.RO] (cit. on pp. 17, 149).

[Huber 2023b] Martin Huber, Sébastien Ourselin, Christos Bergeles, and Tom Vercauteren. "Deep Homography Prediction for Endoscopic Camera Motion Imitation Learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 217–226. DOI: 10.1007/978-3-031-43996-4_21 (cit. on pp. 17, 131).

[Hutzl 2015] Jessica Hutzl, Andreas Bihlmaier, Martin Wagner, Hannes Götz Kenngott, Beat Peter Müller, and Heinz Wörn. "Knowledge-based workspace optimization of a redundant robot for minimally invasive robotic surgery (MIRS)". In: *2015 IEEE International Conference on Robotics and Biomimet-*

ics (ROBIO). 2015, pp. 1403–1408. DOI: 10.1109/ROBIO.2015.7418967 (cit. on p. 49).

[Isaacson 2011] Walter Isaacson. *Steve Jobs: The Exclusive Biography*. Simon & Schuster, 2011 (cit. on p. 5).

[Islam 2019] Mobarakol Islam, Daniel Anojan Atputharuban, Ravikiran Ramesh, and Hongliang Ren. "Real-Time Instrument Segmentation in Robotic Surgery Using Auxiliary Supervised Deep Adversarial Learning". In: *IEEE Robotics and Automation Letters* 4.2 (2019), pp. 2188–2195. DOI: 10.1109/LRA.2019.2900854 (cit. on p. 61).

[Ji 2018] Jessica J. Ji, Sanjay Krishnan, Vatsal Patel, Danyal Fer, and Ken Goldberg. "Learning 2D Surgical Camera Motion From Demonstrations". In: *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*. 2018, pp. 35–42. DOI: 10.1109/COASE.2018.8560468 (cit. on pp. 60, 70).

[Jiang 2023] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL] (cit. on p. 147).

[Jin 2018a] Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. "Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018, pp. 691–699. DOI: 10.1109/WACV.2018.00081 (cit. on p. 61).

[Jin 2018b] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. "SV-RCNet: Workflow Recognition From Surgical Videos

Using Recurrent Convolutional Network". In: *IEEE Transactions on Medical Imaging* 37.5 (2018), pp. 1114–1126. DOI: 10.1109/TMI.2017.2787657 (cit. on p. 62).

[Jin 2020] Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. "Multi-task recurrent convolutional network with correlation loss for surgical video analysis". In: *Medical Image Analysis* 59 (2020), p. 101572. ISSN: 1361-8415. DOI: 10.1016/j.media.2019.101572. URL: https://www.sciencedirect.com/science/article/pii/S1361841519301124 (cit. on p. 62).

[John 2020] Ace St John, Ilaria Caturegli, Natalia S Kubicki, and Stephen M Kavic. "The rise of minimally invasive surgery: 16 year analysis of the progressive replacement of open surgery with laparoscopy". In: *JSLS: Journal of the Society of Laparoscopic & Robotic Surgeons* 24.4 (2020). PMID: 33510568. DOI: 10.4293/JSLS.2020.00076 (cit. on p. 39).

[Kamezaki 2016] Mitsuhiro Kamezaki, Hiroyuki Ishii, Tatsuzo Ishida, Masatoshi Seki, Ken Ichiryu, Yo Kobayashi, Kenji Hashimoto, Shigeki Sugano, Atsuo Takanishi, Masakatsu G. Fujie, Shuji Hashimoto, and Hiroshi Yamakawa. "Design of four-arm four-crawler disaster response robot OCTOPUS". In: *2016 IEEE International Conference on Robotics and Automation (ICRA)* (2016), pp. 2840–2845. DOI: 10.1109/ICRA.2016.7487447 (cit. on p. 78).

[Kane 2020] William J. Kane, Eric J. Charles, J. Hunter Mehaffey, Robert B. Hawkins, Kathleen B. Meneses, Carlos A. Tache-Leon, and Zequan Yang. "Robotic compared with laparoscopic cholecystectomy: A propensity matched analysis". In: *Surgery* 167.2 (2020), pp. 432–435. ISSN: 0039-6060. DOI: 10.1016/j.surg.2019.07.020. URL: https://www.sciencedirect.com/science/article/pii/S0039606019305264 (cit. on p. 48).

[Kassahun 2016] Yohannes Kassahun, Bingbin Yu, Abraham Temesgen Tibebu, Danail Stoyanov, Stamatia Giannarou, Jan Hendrik Metzen, and Emmanuel Vander Poorten. "Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions". In: *International Journal of Computer Assisted Radiology and Surgery* 11.4 (2016), pp. 553–568. DOI: 10.1007/s11548-015-1305-z (cit. on p. 69).

[Kawka 2023] Michal Kawka, Yuman Fong, and Tamara MH Gall. "Laparoscopic versus robotic abdominal and pelvic surgery: a systematic review of randomised controlled trials". In: *Surgical Endoscopy* 37.9 (2023), pp. 6672–6681. DOI: 10.1007/s00464-023-10275-8 (cit. on p. 44).

[Khosla 2021] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. *Supervised Contrastive Learning*. 2021. arXiv: 2004.11362 [cs.LG] (cit. on p. 66).

[King 2013] Brady W. King, Luke A. Reisner, Abhilash K. Pandya, Anthony M. Composto, R. Darin Ellis, and Michael D. Klein. "Towards an Autonomous Robot for Camera Control During Laparoscopic Surgery". In: *Journal of Laparoendoscopic & Advanced Surgical Techniques* 23.12 (2013). PMID: 24195784, pp. 1027–1030. DOI: 10.1089/lap.2013.0304 (cit. on p. 59).

[Kirillov 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. "Segment Anything". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 3992–4003. DOI: 10.1109/ICCV51070.2023.00371 (cit. on pp. 61, 74, 85).

[Kitaguchi 2020] Daichi Kitaguchi, Nobuyoshi Takeshita, Hiroki Matsuzaki, Hiroaki Takano, Yohei Owada, Tsuyoshi

Enomoto, Tatsuya Oda, Hirohisa Miura, Takahiro Yamanashi, Masahiko Watanabe, et al. "Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach". In: *Surgical Endoscopy* 34.11 (2020), pp. 4924–4931. DOI: 10.1007/s00464-019-07281-0 (cit. on p. 62).

[Kitaguchi 2022] Daichi Kitaguchi, Nobuyoshi Takeshita, Hiro Hasegawa, and Masaaki Ito. "Artificial intelligence-based computer vision in surgery: Recent advances and future perspectives". In: *Annals of Gastroenterological Surgery* 6.1 (2022), pp. 29–36. DOI: 10.1002/ags3.12513. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ags3.12513 (cit. on p. 36).

[Koenig 2004] Nathan Koenig and Andrew Howard. "Design and use paradigms for gazebo, an open-source multi-robot simulator". In: *2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)(IEEE Cat. No. 04CH37566)*. Vol. 3. Ieee. 2004, pp. 2149–2154. DOI: 10.1109/IROS.2004.1389727 (cit. on p. 150).

[Kurmann 2017] Thomas Kurmann, Pablo Marquez Neila, Xiaofei Du, Pascal Fua, Danail Stoyanov, Sebastian Wolf, and Raphael Sznitman. "Simultaneous Recognition and Pose Estimation of Instruments in Minimally Invasive Surgery". In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017*. Cham: Springer International Publishing, 2017, pp. 505–513. ISBN: 978-3-319-66185-8. DOI: 10.1007/978-3-319-66185-8_57 (cit. on p. 61).

[Kwon 2020] Hyungju Kwon. "Impact of bedside assistant on outcomes of robotic thyroid surgery: a STROBE-compliant retrospective case-control study". In: *Medicine* 99.36 (2020). PMID: 32899100. DOI: 10.1097/MD.0000000000022133 (cit. on p. 49).

[Labbé 2021] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. "Single-view robot pose and joint

angle estimation via render & compare". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 1654–1663. DOI: 10.1109/CVPR46437.2021.00170 (cit. on pp. 56, 80).

[Laina 2017] Iro Laina, Nicola Rieke, Christian Rupprecht, Josué Page Vizcaíno, Abouzar Eslami, Federico Tombari, and Nassir Navab. "Concurrent Segmentation and Localization for Tracking of Surgical Instruments". In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017*. Cham: Springer International Publishing, 2017, pp. 664–672. ISBN: 978-3-319-66185-8. DOI: 10.1007/978-3-319-66185-8_75 (cit. on p. 61).

[Lalys 2014] Florent Lalys and Pierre Jannin. "Surgical process modelling: a review". In: *International journal of computer assisted radiology and surgery* 9.3 (2014), pp. 495–511. DOI: 10.1007/s11548-013-0940-5 (cit. on p. 62).

[Le 2020] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. "Deep Homography Estimation for Dynamic Scenes". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 7649–7658. DOI: 10.1109/CVPR42600.2020.00767 (cit. on pp. 119, 120).

[LeCun 2022] Yann LeCun. "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27". In: *Open Review* 62.1 (2022). URL: https://openreview.net/forum?id=BZ5a1r-kVsf (cit. on p. 37).

[Lee 2020] Timothy E Lee, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Oliver Kroemer, Dieter Fox, and Stan Birchfield. "Camera-to-robot pose estimation from a single image". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. May 2020, pp. 9426–9432. DOI: 10.1109/ICRA40945.2020.9196596 (cit. on p. 80).

[Li 2020a] Weibing Li, Philip Wai Yan Chiu, and Zheng Li. "An Accelerated Finite-Time Convergent Neural

Network for Visual Servoing of a Flexible Surgical Endoscope With Physical and RCM Constraints". In: *IEEE Transactions on Neural Networks and Learning Systems* 31.12 (2020), pp. 5272–5284. DOI: 10.1109/TNNLS.2020.2965553 (cit. on pp. 60, 102).

[Li 2020b] Yang Li, Florian Richter, Jingpei Lu, Emily K. Funk, Ryan K. Orosco, Jianke Zhu, and Michael C. Yip. "SuPer: A Surgical Perception Framework for Endoscopic Tissue Manipulation With Surgical Robotics". In: *IEEE Robotics and Automation Letters* 5.2 (Apr. 2020), pp. 2294–2301. ISSN: 2377-3774. DOI: 10.1109/lra.2020.2970659 (cit. on p. 62).

[Li 2021a] Bin Li, Bo Lu, Yiang Lu, Qi Dou, and Yun-Hui Liu. "Data-driven Holistic Framework for Automated Laparoscope Optimal View Control with Learning-based Depth Perception". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. 2021, pp. 12366–12372. DOI: 10.1109/ICRA48506.2021.9562083 (cit. on p. 132).

[Li 2021b] Ling Li, Xiaojian Li, Shanlin Yang, Shuai Ding, Alireza Jolfaei, and Xi Zheng. "Unsupervised-Learning-Based Continuous Depth and Motion Estimation With Monocular Endoscopy for Virtual Reality Minimally Invasive Surgery". In: *IEEE Transactions on Industrial Informatics* 17.6 (2021), pp. 3920–3928. DOI: 10.1109/TII.2020.3011067 (cit. on p. 62).

[Li 2022a] Bin Li, Bo Lu, Ziyi Wang, Fangxun Zhong, Qi Dou, and Yun-Hui Liu. "Learning Laparoscope Actions via Video Features for Proactive Robotic Field-of-View Control". In: *IEEE Robotics and Automation Letters* 7.3 (2022), pp. 6653–6660. DOI: 10.1109/LRA.2022.3173442 (cit. on p. 132).

[Li 2022b] Bin Li, Ruofeng Wei, Jiaqi Xu, Bo Lu, Chi Hang Yee, Chi Fai Ng, Pheng-Ann Heng, Qi Dou, and Yun-Hui Liu. "3D Perception based Imitation Learn-

ing under Limited Demonstration for Laparoscope Control in Robotic Surgery". In: *2022 International Conference on Robotics and Automation (ICRA)*. 2022, pp. 7664–7670. DOI: 10.1109/ICRA46639.2022.9812010 (cit. on p. 132).

[Li 2022c] Wenda Li, Yuichiro Hayashi, Masahiro Oda, Takayuki Kitasaka, Kazunari Misawa, and Kensaku Mori. "Geometric Constraints for Self-supervised Monocular Depth Estimation on Laparoscopic Images with Dual-task Consistency". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Cham: Springer Nature Switzerland, 2022, pp. 467–477. ISBN: 978-3-031-16440-8. DOI: 10.1007/978-3-031-16440-8_45 (cit. on p. 62).

[Li 2023a] Ling Li, Xiaojian Li, Bo Ouyang, Hangjie Mo, Hongliang Ren, and Shanlin Yang. "Three-Dimensional Collision Avoidance Method for Robot-Assisted Minimally Invasive Surgery". In: *Cyborg and Bionic Systems* 4 (2023), p. 0042. DOI: 10.34133/cbsystems.0042 (cit. on p. 49).

[Li 2023b] Wenda Li, Yuichiro Hayashi, Masahiro Oda, Takayuki Kitasaka, Kazunari Misawa, and Kensaku Mori. "Multi-view Guidance for Self-supervised Monocular Depth Estimation on Laparoscopic Images via Spatio-Temporal Correspondence". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Cham: Springer Nature Switzerland, 2023, pp. 429–439. ISBN: 978-3-031-43996-4. DOI: 10.1007/978-3-031-43996-4_41 (cit. on p. 62).

[Liu 2015] May Liu and Myriam Curet. "A Review of Training Research and Virtual Reality Simulators for the da Vinci Surgical System". In: *Teaching and Learning in Medicine* 27.1 (2015). PMID: 25584468, pp. 12–26. DOI: 10.1080/10401334.2014.979181 (cit. on p. 78).

[Liu 2018] YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. "Imitation from Observation: Learn-

ing to Imitate Behaviors from Raw Video via Context Translation". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2018, pp. 1118–1125. DOI: 10.1109/ICRA.2018.8462901 (cit. on p. 64).

[Liu 2022] Siqi Liu, Guy Lever, Zhe Wang, Josh Merel, S. M. Ali Eslami, Daniel Hennes, Wojciech M. Czarnecki, Yuval Tassa, Shayegan Omidshafiei, Abbas Abdolmaleki, Noah Y. Siegel, Leonard Hasenclever, Luke Marris, Saran Tunyasuvunakool, H. Francis Song, Markus Wulfmeier, Paul Muller, Tuomas Haarnoja, Brendan Tracey, Karl Tuyls, Thore Graepel, and Nicolas Heess. "From motor control to team play in simulated humanoid football". In: *Science Robotics* 7.69 (2022), eabo0235. DOI: 10.1126/scirobotics.abo0235. URL: https://www.science.org/doi/abs/10.1126/scirobotics.abo0235 (cit. on p. 63).

[Lou 2024] Ange Lou and Jack Noble. *WS-SfMLearner: Self-supervised Monocular Depth and Ego-motion Estimation on Surgical Videos with Unknown Camera Parameters*. 2024. arXiv: 2308.11776 [cs.CV] (cit. on p. 62).

[Lowe 2004] David G Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision* 60.2 (2004), pp. 91–110. DOI: doi.org/10.1023/B:VISI.0000029664.99615.94 (cit. on p. 108).

[Lu 2023] Jingpei Lu, Florian Richter, and Michael C Yip. "Markerless camera-to-robot pose estimation via self-supervised sim-to-real transfer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 21296–21306. DOI: 10.1109/CVPR52729.2023.02040 (cit. on pp. 56, 80).

[Luengo 2022] Imanol Luengo, Maria Grammatikopoulou, Rahim Mohammadi, Chris Walsh, Chinedu Innocent Nwoye, Deepak Alapatt, Nicolas Padoy, Zhen-Liang Ni,

Chen-Chen Fan, Gui-Bin Bian, Zeng-Guang Hou, Heonjin Ha, Jiacheng Wang, Haojie Wang, Dong Guo, Lu Wang, Guotai Wang, Mobarakol Islam, Bharat Giddwani, Ren Hongliang, Theodoros Pissas, Claudio Ravasio, Martin Huber, Jeremy Birch, Joan M. Nunez Do Rio, Lyndon da Cruz, Christos Bergeles, Hongyu Chen, Fucang Jia, Nikhil KumarTomar, Debesh Jha, Michael A. Riegler, Pal Halvorsen, Sophia Bano, Uddhav Vaghela, Jianyuan Hong, Haili Ye, Feihong Huang, Da-Han Wang, and Danail Stoyanov. *2020 CATARACTS Semantic Segmentation Challenge*. 2022. arXiv: 2110.10965 [eess.IV] (cit. on p. 18).

[Lynch 2019] Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. *Learning Latent Plans from Play*. 2019. arXiv: 1903.01973 [cs.RO] (cit. on p. 65).

[Ma 2014] Songde Ma and Zhanyi Hu. "Hand-Eye Calibration". In: *Computer Vision: A Reference Guide*. Ed. by Katsushi Ikeuchi. Boston, MA: Springer US, 2014, pp. 355–358. ISBN: 978-0-387-31439-6. DOI: 10.1007/978-0-387-31439-6_168. URL: https://doi.org/10.1007/978-0-387-31439-6_168 (cit. on p. 53).

[Ma 2019] Xin Ma, Chengzhi Song, Philip Waiyan Chiu, and Zheng Li. "Autonomous Flexible Endoscope for Minimally Invasive Surgery With Enhanced Safety". In: *IEEE Robotics and Automation Letters* 4.3 (2019), pp. 2607–2613. DOI: 10.1109/LRA.2019.2895273 (cit. on pp. 59, 60, 102).

[Ma 2020] Xin Ma, Chengzhi Song, Philip Waiyan Chiu, and Zheng Li. "Visual Servo of a 6-DOF Robotic Stereo Flexible Endoscope Based on da Vinci Research Kit (dVRK) System". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 820–827. DOI: 10.1109/LRA.2020.2965863 (cit. on pp. 59, 60, 102).

[Macenski 2022] Steven Macenski, Tully Foote, Brian Gerkey, Chris Lalancette, and William Woodall. "Robot Operat-

ing System 2: Design, architecture, and uses in the wild". In: *Science robotics* 7.66 (2022), eabm6074. DOI: 10.1126/scirobotics.abm6074 (cit. on p. 150).

[Macrotrends 2024] Macrotrends. *Intuitive Surgical Financial Statements 2009-2022*. https://www.macrotrends.net/stocks/charts/ISRG/intuitive-surgical/financial-statements. Accessed: 2024-02-14. 2024 (cit. on p. 45).

[Magyar 2023] Bence Magyar, Denis Stogl, Karsten Knese, and Community. *Generic and simple controls framework for ROS 2*. https://github.com/ros-controls/ros2_control. 2023 (cit. on p. 151).

[Mahony 2002] Robert Mahony and Jonathan H Manton. "The geometry of the Newton method on non-compact Lie groups". In: *Journal of Global Optimization* 23.3-4 (2002), pp. 309–327. DOI: 10.1023/A:1016586831090 (cit. on p. 87).

[Maier-Hein 2021] Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M. Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Anna Kisilenko, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Minu Tizabi, Matthias Eisenmann, Tim J. Adler, Janek Gröhl, Melanie Schellenberg, Silvia Seidlitz, T. Y. Emmy Lai, Bünyamin Pekdemir, Veith Roethlingshoefer, Fabian Both, Sebastian Bittel, Marc Mengler, Lars Mündermann, Martin Apitz, Annette Kopp-Schneider, Stefanie Speidel, Hannes G. Kenngott, and Beat P. Müller-Stich. *Heidelberg Colorectal Data Set for Surgical Data Science in the Sensor Operating Room*. 2021. arXiv: 2005.03501 [cs.CV] (cit. on pp. 69, 75, 76, 120).

[Maier-Hein 2022] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, Hirenkumar Nakawala, Adrian Park, Carla Pugh, Danail Stoyanov, Swaroop S. Vedula, Kevin Cleary, Gabor Fichtinger, Germain

Forestier, Bernard Gibaud, Teodor Grantcharov, Makoto Hashizume, Doreen Heckmann-Nötzel, Hannes G. Kenngott, Ron Kikinis, Lars Mündermann, Nassir Navab, Sinan Onogur, Tobias Roß, Raphael Sznitman, Russell H. Taylor, Minu D. Tizabi, Martin Wagner, Gregory D. Hager, Thomas Neumuth, Nicolas Padoy, Justin Collins, Ines Gockel, Jan Goedeke, Daniel A. Hashimoto, Luc Joyeux, Kyle Lam, Daniel R. Leff, Amin Madani, Hani J. Marcus, Ozanan Meireles, Alexander Seitel, Dogu Teber, Frank Ückert, Beat P. Müller-Stich, Pierre Jannin, and Stefanie Speidel. "Surgical data science – from concepts toward clinical translation". In: *Medical Image Analysis* 76 (2022), p. 102306. ISSN: 1361-8415. DOI: 10.1016/j.media.2021.102306. URL: https://www.sciencedirect.com/science/article/pii/S1361841521003510 (cit. on p. 69).

[Majumdar 1993] Sisir Kumar Majumdar. "A short history of gastrointestinal endoscopy." In: *Bulletin of the Indian Institute of History of Medicine (Hyderabad)* 23.1 (1993). PMID: 11639385, pp. 67–86 (cit. on p. 38).

[Majumder 2020] Arnab Majumder, Maria S. Altieri, and L. Michael Brunt. "How do I do it: laparoscopic cholecystectomy". In: *Annals of Laparoscopic and Endoscopic Surgery* 5.0 (2020). ISSN: 2518-6973. DOI: 10.21037/ales.2020.02.06. URL: https://ales.amegroups.org/article/view/5766 (cit. on pp. 40–42).

[Malis 2007] Ezio Malis and Manuel Vargas. "Deeper understanding of the homography decomposition for vision-based control". PhD thesis. INRIA, 2007 (cit. on p. 72).

[Malpani 2016] Anand Malpani, Colin Lea, Chi Chiung Grace Chen, and Gregory D Hager. "System events: readily accessible features for surgical phase detection". In: *International journal of computer assisted radiology and surgery* 11.6 (2016), pp. 1201–1209. DOI: 10.1007/s11548-016-1409-0 (cit. on p. 62).

[Mariani 2020] Andrea Mariani, Giorgia Colaci, Tommaso Da Col, Nicole Sanna, Eleonora Vendrame, Arianna Menciassi, and Elena De Momi. "An Experimental Comparison Towards Autonomous Camera Navigation to Optimize Training in Robot Assisted Surgery". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 1461–1467. DOI: 10.1109/LRA.2020.2965067 (cit. on p. 59).

[Marzullo 2021] Aldo Marzullo, Sara Moccia, Michele Catellani, Francesco Calimeri, and Elena De Momi. "Towards realistic laparoscopic image generation using image-domain translation". In: *Computer Methods and Programs in Biomedicine* 200 (2021), p. 105834. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2020.105834. URL: https://www.sciencedirect.com/science/article/pii/S0169260720316679 (cit. on p. 64).

[Mathieu 2016] Michael Mathieu, Camille Couprie, and Yann LeCun. *Deep multi-scale video prediction beyond mean square error*. 2016. arXiv: 1511.05440 [cs.LG] (cit. on p. 66).

[Maynou 2021] Laia Maynou, Winta T Mehtsun, Victoria Serra-Sastre, and Irene Papanicolas. "Patterns of adoption of robotic radical prostatectomy in the United States and England". In: *Health services research* 56 (2021). PMID: 34350592, pp. 1441–1461. DOI: 10.1111/1475-6773.13706 (cit. on p. 45).

[Mischinger 2020] Hans-Jörg Mischinger, Doris Wagner, Peter Kornprat, Heinz Bacher, and Georg Werkgartner. "The "critical view of safety (CVS)" cannot be applied—What to do? Strategies to avoid bile duct injuries". In: *European Surgery* 53 (Sept. 2020), pp. 1–7. DOI: 10.1007/s10353-020-00660-1 (cit. on pp. 41, 42).

[Mitsinikos 2017] Emmanuel Mitsinikos, George A Abdelsayed, Zoe Bider, Patrick S Kilday, Peter A Elliott, Pooya Banapour, and Gary W Chien. "Does the level of assistant experience impact operative outcomes for

robot-assisted partial nephrectomy?" In: *Journal of endourology* 31.1 (2017), pp. 38–42. DOI: 10.1089/end.2016.0508 (cit. on p. 49).

[Mitsuishi 2013] Mamoru Mitsuishi, Akio Morita, Naohiko Sugita, Shigeo Sora, Ryo Mochizuki, Keiji Tanimoto, Young Min Baek, Hiroki Takahashi, and Kanako Harada. "Master–slave robotic platform and its feasibility study for micro-neurosurgery". In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 9.2 (2013). PMID: 22588785, pp. 180–189. DOI: 10.1002/rcs.1434 (cit. on pp. 70, 76).

[Mnih 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. *Playing Atari with Deep Reinforcement Learning*. 2013. arXiv: 1312.5602 [cs.LG] (cit. on p. 63).

[Monfared 2022] Sara Monfared, Dimitrios I Athanasiadis, Luke Umana, Edward Hernandez, Hamed Asadi, Cameron L Colgate, Denny Yu, and Dimitrios Stefanidis. "A comparison of laparoscopic and robotic ergonomic risk". In: *Surgical Endoscopy* 36.11 (2022), pp. 8397–8402. DOI: 10.1007/s00464-022-09105-0 (cit. on p. 45).

[Mountney 2010] Peter Mountney, Danail Stoyanov, and Guang-Zhong Yang. "Three-Dimensional Tissue Deformation Recovery and Tracking". In: *IEEE Signal Processing Magazine* 27.4 (2010), pp. 14–24. DOI: 10.1109/MSP.2010.936728 (cit. on pp. 76, 120).

[Mower 2023a] Christopher E Mower, Martin Huber, Huanyu Tian, Ayoob Davoodi, Emmanuel Vander Poorten, Tom Vercauteren, and Christos Bergeles. "Vision and Contact based Optimal Control for Autonomous Trocar Docking". In: *12th Conference on New Technologies for Computer and Robot Assisted Surgery*. 2023 (cit. on p. 18).

[Mower 2023b] Christopher E. Mower, João Moura, Nazanin Zamani Behabadi, Sethu Vijayakumar, Tom Vercauteren, and Christos Bergeles. "OpTaS: An Optimization-

based Task Specification Library for Trajectory Optimization and Model Predictive Control". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2023. DOI: 10.1109/icra48891.2023.10161272 (cit. on p. 94).

[Münzer 2013] Bernd Münzer, Klaus Schoeffmann, and Laszlo Böszörmenyi. "Detection of circular content area in endoscopic videos". In: *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. 2013, pp. 534–536. DOI: 10.1109/CBMS.2013.6627865 (cit. on p. 121).

[Nair 2017] Ashvin Nair, Dian Chen, Pulkit Agrawal, Phillip Isola, Pieter Abbeel, Jitendra Malik, and Sergey Levine. "Combining self-supervised learning and imitation for vision-based rope manipulation". In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 2017, pp. 2146–2153. DOI: 10.1109/ICRA.2017.7989247 (cit. on p. 65).

[Ng 2000] Andrew Y Ng, Stuart J Russell, et al. "Algorithms for inverse reinforcement learning." In: *Icml*. Vol. 1. 2000, p. 2 (cit. on p. 64).

[Nguyen 2018] Ty Nguyen, Steven W. Chen, Shreyas S. Shivakumar, Camillo Jose Taylor, and Vijay Kumar. "Unsupervised Deep Homography: A Fast and Robust Homography Estimation Model". In: *IEEE Robotics and Automation Letters* 3.3 (2018), pp. 2346–2353. DOI: 10.1109/LRA.2018.2809549 (cit. on p. 119).

[Niemeyer 2008] Günter Niemeyer, Carsten Preusche, and Gerd Hirzinger. "Telerobotics". In: *Springer Handbook of Robotics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 741–757. ISBN: 978-3-540-30301-5. DOI: 10.1007/978-3-540-30301-5_32 (cit. on p. 78).

[Nowruzi 2017] Farzan Erlik Nowruzi, Robert Laganiere, and Nathalie Japkowicz. "Homography Estimation from Image Pairs with Hierarchical Convolutional Networks". In: *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2017, pp. 904–

911. DOI: 10.1109/ICCVW.2017.111 (cit. on p. 119).

[Nwoye 2021] Chinedu Innocent Nwoye. "Deep learning methods for the detection and recognition of surgical tools and activities in laparoscopic videos". PhD thesis. Université de Strasbourg, 2021 (cit. on p. 62).

[OED Online 2023] OED Online. *endoscopy (n.)* OED Online. 2023 (cit. on p. 38).

[Olson 2011] Edwin Olson. "AprilTag: A robust and flexible visual fiducial system". In: *2011 IEEE International Conference on Robotics and Automation*. 2011, pp. 3400–3407. DOI: 10.1109/ICRA.2011.5979561 (cit. on p. 80).

[Omote 1999] Kazuhiko Omote, Hubertus Feussner, Andreas Ungeheuer, Klaus Arbter, Guo-Qing Wei, J.Rüdiger Siewert, and Gerd Hirzinger. "Self-guided robotic camera control for laparoscopic surgery compared with human camera control". In: *The American Journal of Surgery* 177.4 (1999), pp. 321–324. ISSN: 0002-9610. DOI: 10.1016/S0002-9610(99)00055-0. URL: https://www.sciencedirect.com/science/article/pii/S0002961099000550 (cit. on p. 59).

[OpenAI 2024] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL] (cit. on p. 147).

[Oquab 2024] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. *DINOv2: Learning Robust Visual Features without Supervision*. 2024. arXiv: 2304.07193 [cs.CV] (cit. on p. 61).

[Osa 2010] Takayuki Osa, Christoph Staub, and Alois Knoll. "Framework of automatic robot surgery system using Visual servoing". In: *2010 IEEE/RSJ Interna-*

tional Conference on Intelligent Robots and Systems. 2010, pp. 1837–1842. DOI: 10.1109/IROS.2010.5650301 (cit. on p. 102).

[Osa 2018] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. "An Algorithmic Perspective on Imitation Learning". In: *Foundations and Trends in Robotics* 7.1–2 (2018), pp. 1–179. ISSN: 1935-8261. DOI: 10.1561/2300000053 (cit. on p. 64).

[Pachtrachai 2016] Krittin Pachtrachai, Max Allan, Vijay Pawar, Stephen Hailes, and Danail Stoyanov. "Hand-eye calibration for robotic assisted minimally invasive surgery without a calibration object". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2016, pp. 2485–2491. DOI: 10.1109/IROS.2016.7759387 (cit. on p. 80).

[Padoy 2019] Nicolas Padoy. "Machine and deep learning for workflow recognition during surgery". In: *Minimally Invasive Therapy & Allied Technologies* 28.2 (2019). PMID: 30849261, pp. 82–90. DOI: 10.1080/13645706.2019.1584116 (cit. on p. 62).

[Pakhomov 2019] Daniil Pakhomov, Vittal Premachandran, Max Allan, Mahdi Azizian, and Nassir Navab. "Deep Residual Learning for Instrument Segmentation in Robotic Surgery". In: *Machine Learning in Medical Imaging*. Cham: Springer International Publishing, 2019, pp. 566–573. ISBN: 978-3-030-32692-0. DOI: 10.1007/978-3-030-32692-0_65 (cit. on p. 61).

[Pandya 2014] Abhilash Pandya, Luke A. Reisner, Brady King, Nathan Lucas, Anthony Composto, Michael Klein, and Richard Darin Ellis. "A Review of Camera Viewpoint Automation in Robotic and Laparoscopic Surgery". In: *Robotics* 3.3 (2014), pp. 310–329. ISSN: 2218-6581. DOI: 10.3390/robotics3030310. URL: https://www.mdpi.com/2218-6581/3/3/310 (cit. on p. 58).

[Park 1994] F.C. Park and B.J. Martin. "Robot sensor calibration: solving AX=XB on the Euclidean group".

In: *IEEE Transactions on Robotics and Automation* 10.5 (1994), pp. 717–721. DOI: 10.1109/70.326576 (cit. on p. 55).

[Patel 2021] Sejal Patel, Maroeska M Rovers, Michiel JP Sedelaar, Petra LM Zusterzeel, Ad FTM Verhagen, Camiel Rosman, and Janneke PC Grutters. "How can robot-assisted surgery provide value for money?" In: *BMJ Surgery, Interventions, & Health Technologies* 3.1 (2021). DOI: 10.1136/bmjsit-2020-000042 (cit. on p. 47).

[Patel 2023] Nainita Patel, Kamlesh Chaudhari, Garapati Jyotsna, and Jalormy S Joshi. "Surgical Frontiers: a comparative review of robotics versus laparoscopy in gynecological interventions". In: *Cureus* 15.11 (2023). DOI: 10.7759/cureus.49752 (cit. on p. 48).

[Pathak 2018] Deepak Pathak, Parsa Mahmoudieh, Guanghao Luo, Pulkit Agrawal, Dian Chen, Fred Shentu, Evan Shelhamer, Jitendra Malik, Alexei A. Efros, and Trevor Darrell. "Zero-Shot Visual Imitation". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018, pp. 2131–21313. DOI: 10.1109/CVPRW.2018.00278 (cit. on p. 65).

[Pomerleau 1991] Dean A. Pomerleau. "Efficient Training of Artificial Neural Networks for Autonomous Navigation". In: *Neural Computation* 3.1 (1991), pp. 88–97. DOI: 10.1162/neco.1991.3.1.88 (cit. on p. 64).

[Ravi 2020] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. "Accelerating 3D Deep Learning with PyTorch3D". In: *arXiv:2007.08501* (2020) (cit. on p. 80).

[Rivoir 2019] Dominik Rivoir, Sebastian Bodenstedt, Felix von Bechtolsheim, Marius Distler, Jürgen Weitz, and Stefanie Speidel. "Unsupervised Temporal Video Segmentation as an Auxiliary Task for Predicting the Remaining Surgery Duration". In: *OR 2.0 Context-Aware Operating Theaters and Machine*

Learning in Clinical Neuroimaging. Cham: Springer International Publishing, 2019, pp. 29–37. ISBN: 978-3-030-32695-1. DOI: 10.1007/978-3-030-32695-1_4 (cit. on p. 62).

[Rombach 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10674–10685. DOI: 10.1109/CVPR52688.2022.01042 (cit. on p. 146).

[Ross 2018] Tobias Ross, David Zimmerer, Anant Vemuri, Fabian Isensee, Manuel Wiesenfarth, Sebastian Bodenstedt, Fabian Both, Philip Kessler, Martin Wagner, Beat Müller, et al. "Exploiting the potential of unlabeled endoscopic video data with self-supervised learning". In: *International journal of computer assisted radiology and surgery* 13.6 (2018), pp. 925–933. DOI: 10.1007/s11548-018-1772-0 (cit. on p. 62).

[Ross 2020] Tobias Ross, Annika Reinke, Peter M. Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Pablo Arbeláez, Gui-Bin Bian, Sebastian Bodenstedt, Jon Lindström Bolmgren, Laura Bravo-Sánchez, Hua-Bin Chen, Cristina González, Dong Guo, Pål Halvorsen, Pheng-Ann Heng, Enes Hosgor, Zeng-Guang Hou, Fabian Isensee, Debesh Jha, Tingting Jiang, Yueming Jin, Kadir Kirtac, Sabrina Kletz, Stefan Leger, Zhixuan Li, Klaus H. Maier-Hein, Zhen-Liang Ni, Michael A. Riegler, Klaus Schoeffmann, Ruohua Shi, Stefanie Speidel, Michael Stenzel, Isabell Twick, Gutai Wang, Jiacheng Wang, Liansheng Wang, Lu Wang, Yujie Zhang, Yan-Jie Zhou, Lei Zhu, Manuel Wiesenfarth, Annette Kopp-Schneider, Beat P. Müller-Stich, and Lena Maier-Hein. *Robust Medical In-*

strument Segmentation Challenge 2019. 2020. arXiv: 2003.10299 [cs.CV] (cit. on p. 76).

[Rudin 2022] Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. *Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning*. 2022. arXiv: 2109.11978 [cs.RO] (cit. on p. 63).

[Rusinkiewicz 2001] S. Rusinkiewicz and M. Levoy. "Efficient variants of the ICP algorithm". In: *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*. 2001, pp. 145–152. DOI: 10.1109/IM.2001.924423 (cit. on p. 87).

[Saeidi 2022] H. Saeidi, J. D. Opfermann, M. Kam, S. Wei, S. Leonard, M. H. Hsieh, J. U. Kang, and A. Krieger. "Autonomous robotic laparoscopic surgery for intestinal anastomosis". In: *Science Robotics* 7.62 (2022). DOI: 10.1126/scirobotics.abj2908 (cit. on p. 78).

[Safeea 2018] Mohammad Safeea and Pedro Neto. "Kuka sunrise toolbox: Interfacing collaborative robots with matlab". In: *IEEE Robotics & Automation Magazine* 26.1 (2018), pp. 91–96. DOI: 10.1109/MRA.2018.2877776 (cit. on p. 151).

[SAGES 2010] SAGES. *Basic room setup*. https://www.sages.org/image-library/basic-room-setup/. Fundamentals: Laparoscopy-General Principals, Accessed: 2024-01-26. 2010 (cit. on p. 40).

[Sandoval 2021] J. Sandoval, M.A. Laribi, J.P. Faure, C. Brèque, J.P. Richer, and S. Zeghloul. "Towards an Autonomous Robot-Assistant for Laparoscopy Using Exteroceptive Sensors: Feasibility Study and Implementation". In: *IEEE Robotics and Automation Letters* 6.4 (2021), pp. 6473–6480. DOI: 10.1109/LRA.2021.3094644 (cit. on p. 60).

[Sarikaya 2017] Duygu Sarikaya, Jason J. Corso, and Khurshid A. Guru. "Detection and Localization of Robotic Tools in Robot-Assisted Surgery Videos Using Deep Neural Networks for Region Proposal and Detection". In: *IEEE Transactions on Medical Imaging*

36.7 (2017), pp. 1542–1549. DOI: 10.1109/TMI.2017.2665671 (cit. on p. 61).

[Scheikl 2023] Paul Maria Scheikl, Balázs Gyenes, Rayan Younis, Christoph Haas, Gerhard Neumann, Martin Wagner, and Franziska Mathis-Ullrich. *LapGym – An Open Source Framework for Reinforcement Learning in Robot-Assisted Laparoscopic Surgery*. 2023. arXiv: 2302.09606 [cs.RO] (cit. on p. 63).

[Schreiber 2010] Günter Schreiber, Andreas Stemmer, and Rainer Bischoff. "The fast research interface for the kuka lightweight robot". In: *IEEE workshop on innovative robot control architectures for demanding (Research) applications how to modify and enhance commercial controllers (ICRA 2010)*. 2010, pp. 15–21 (cit. on pp. 109, 150).

[Schroff 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A unified embedding for face recognition and clustering". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682 (cit. on p. 66).

[Schulman 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. *Proximal Policy Optimization Algorithms*. 2017. arXiv: 1707.06347 [cs.LG] (cit. on p. 63).

[Seenivasan 2022] Lalithkumar Seenivasan, Mobarakol Islam, Adithya K. Krishna, and Hongliang Ren. "Surgical-VQA: Visual Question Answering in Surgical Scenes Using Transformer". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Cham: Springer Nature Switzerland, 2022, pp. 33–43. ISBN: 978-3-031-16449-1. DOI: 10.1007/978-3-031-16449-1_4 (cit. on p. 147).

[Sermanet 2017] Pierre Sermanet, Kelvin Xu, and Sergey Levine. *Unsupervised Perceptual Rewards for Imitation Learning*. 2017. arXiv: 1612.06699 [cs.CV] (cit. on p. 66).

[Sermanet 2018] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. "Time-Contrastive Networks: Self-Supervised Learning from Video". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2018, pp. 1134–1141. DOI: 10.1109/ICRA.2018.8462891 (cit. on p. 66).

[Serrano-Muñoz 2023] Antonio Serrano-Muñoz, Íñigo Elguea-Aguinaco, Dimitris Chrysostomou, Simon BØgh, and Nestor Arana-Arexolaleiba. "A scalable and unified multi-control framework for KUKA LBR iiwa collaborative robots". In: *2023 IEEE/SICE International Symposium on System Integration (SII)*. IEEE. 2023, pp. 1–5. DOI: 10.1109/SII55687.2023.10039308 (cit. on p. 151).

[Shao 2022] Shuwei Shao, Zhongcai Pei, Weihai Chen, Wentao Zhu, Xingming Wu, Dianmin Sun, and Baochang Zhang. "Self-Supervised monocular depth and ego-Motion estimation in endoscopy: Appearance flow to the rescue". In: *Medical Image Analysis* 77 (2022), p. 102338. ISSN: 1361-8415. DOI: 10.1016/j.media.2021.102338. URL: https://www.sciencedirect.com/science/article/pii/S1361841521003832 (cit. on p. 62).

[Sheetz 2020] Kyle H. Sheetz, Jake Claflin, and Justin B. Dimick. "Trends in the Adoption of Robotic Surgery for Common Surgical Procedures". In: *JAMA Network Open* 3.1 (Jan. 2020), e1918911–e1918911. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2019.18911 (cit. on pp. 37, 38, 44, 50).

[Shi 2024] Zhen Shi. "Laparoscopic vs. open surgery: A comparative analysis of wound infection rates and recovery outcomes". In: *International Wound Journal* 21.3 (2024), e14474. DOI: 10.1111/iwj.14474. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/iwj.14474 (cit. on p. 38).

[Shvets 2018] Alexey A. Shvets, Alexander Rakhlin, Alexandr A. Kalinin, and Vladimir I. Iglovikov. "Automatic In-

strument Segmentation in Robot-Assisted Surgery using Deep Learning". In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2018, pp. 624–628. DOI: 10.1109/ICMLA.2018.00100 (cit. on p. 61).

[Silver 2016] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587 (2016), pp. 484–489. DOI: 10.1038/nature16961 (cit. on p. 63).

[Silver 2017] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. *Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm*. 2017. arXiv: 1712.01815 [cs.AI] (cit. on p. 63).

[Simonyan 2015] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV] (cit. on p. 119).

[Srinivas 2018] Aravind Srinivas, Allan Jabri, Pieter Abbeel, Sergey Levine, and Chelsea Finn. "Universal Planning Networks: Learning Generalizable Representations for Visuomotor Control". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 4732–4741. URL: https://proceedings.mlr.press/v80/srinivas18b.html (cit. on pp. 65, 66).

[Srivastava 2016] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. *Unsupervised Learning of Video Representations using LSTMs*. 2016. arXiv: 1502.04681 [cs.LG] (cit. on p. 66).

[Stauder 2014] Ralf Stauder, Aslı Okur, Loïc Peter, Armin Schneider, Michael Kranzfelder, Hubertus Feussner, and Nassir Navab. "Random Forests for Phase Detection in Surgical Workflow Analysis". In: *Information Processing in Computer-Assisted Interventions*. Cham: Springer International Publishing, 2014, pp. 148–157. ISBN: 978-3-319-07521-1. DOI: 10.1007/978-3-319-07521-1_16 (cit. on p. 62).

[Strobl 2006] Klaus H. Strobl and Gerd Hirzinger. "Optimal Hand-Eye Calibration". In: *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2006), pp. 4647–4653. DOI: 10.1109/IROS.2006.282250 (cit. on p. 80).

[Stucky 2018] Chee-Chee H Stucky, Kate D Cromwell, Rachel K Voss, Yi-Ju Chiang, Karin Woodman, Jeffrey E Lee, and Janice N Cormier. "Surgeon symptoms, strain, and selections: systematic review and meta-analysis of surgical ergonomics". In: *Annals of Medicine and Surgery* 27 (2018), pp. 1–8. DOI: 10.1016/j.amsu.2017.12.013 (cit. on p. 45).

[Su 2021] Yun-Hsuan Su, Kevin Huang, and Blake Hannaford. "Multicamera 3D Viewpoint Adjustment for Robotic Surgery via Deep Reinforcement Learning". In: *Journal of Medical Robotics Research* 06.01n02 (2021), p. 2140003. DOI: 10.1142/S2424905X21400031 (cit. on p. 63).

[Suarez-Ahedo 2023] Carlos Suarez-Ahedo, Alberto Lopez-Reyes, Carlos Martinez-Armenta, Laura E Martinez-Gomez, Gabriela A Martinez-Nava, Carlos Pineda, David R Vanegas-Contla, and Benjamin Domb. "Revolutionizing orthopedics: a comprehensive review of robot-assisted surgery, clinical outcomes, and the future of patient care". In: *Journal of Robotic Surgery* 17.6 (2023), pp. 2575–2581. DOI: 10.1007/s11701-023-01697-6 (cit. on p. 51).

[Sun 2021] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou. "LoFTR: Detector-Free Local Feature Matching with Transformers". In: *2021 IEEE/CVF Confer-*

ence on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, June 2021, pp. 8918–8927. DOI: 10.1109/CVPR46437.2021.00881. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00881 (cit. on p. 136).

[Suvorov 2021] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. *Resolution-robust Large Mask Inpainting with Fourier Convolutions*. 2021. arXiv: 2109.07161 [cs.CV] (cit. on p. 145).

[Suzuki 1970] Shunryu Suzuki. *Zen Mind, Beginner's Mind. Informal Talks on Zen Meditation and Practice*. Weatherhill, New York; Tokyo, 1970 (cit. on p. 148).

[Taniguchi 2010] Kazuhiro Taniguchi, Atsushi Nishikawa, Mitsugu Sekimoto, Takeharu Kobayashi, Kouhei Kazuhara, Takaharu Ichihara, Naoto Kurashita, Shuji Takiguchi, Yuichiro Doki, Masaki Mori, and Fumio Miyazaki. "Classification, Design and Evaluation of Endoscope Robots". In: *Robot Surgery*. Rijeka: IntechOpen, 2010. Chap. 1. DOI: 10.5772/6893 (cit. on p. 58).

[Thapar 2023] Vinaykumar B Thapar, Pinky M Thapar, Ramen Goel, Ramesh Agarwalla, Prashant H Salvi, Amrit M Nasta, and Kamal Mahawar. "Evaluation of 30-day morbidity and mortality of laparoscopic cholecystectomy: a multicenter prospective observational Indian Association of Gastrointestinal Endoscopic Surgeons (IAGES) Study". In: *Surgical endoscopy* 37.4 (2023), pp. 2611–2625. DOI: doi.org/10.1007/s00464-022-09659-z (cit. on p. 44).

[Tian 2024] Huanyu Tian, Martin Huber, Christopher E. Mower, Zhe Han, Changsheng Li, Xingguang Duan, and Christos Bergeles. *Excitation Trajectory Optimization for Dynamic Parameter Identification Using Virtual Constraints in Hands-on Robotic System*. 2024. arXiv: 2401.16566 [cs.RO] (cit. on p. 18).

[Torabi 2018] Faraz Torabi, Garrett Warnell, and Peter Stone. *Behavioral Cloning from Observation*. 2018. arXiv: 1805.01954 [cs.AI] (cit. on p. 65).

[Torabi 2019] Faraz Torabi, Garrett Warnell, and Peter Stone. *Generative Adversarial Imitation from Observation*. 2019. arXiv: 1807.06158 [cs.LG] (cit. on p. 65).

[Touvron 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL] (cit. on p. 147).

[Triggs 2000] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. "Bundle Adjustment — A Modern Synthesis". In: *Vision Algorithms: Theory and Practice*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 298–372. ISBN: 978-3-540-44480-0. DOI: 10.1007/3-540-44480-7_21 (cit. on p. 90).

[Tsai 1989] R.Y. Tsai and R.K. Lenz. "A new technique for fully autonomous and efficient 3D robotics hand/eye calibration". In: *IEEE Transactions on Robotics and Automation* 5.3 (1989), pp. 345–358. DOI: 10.1109/70.34770 (cit. on p. 55).

[Twinanda 2017] Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. "EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos". In: *IEEE Transactions on Medical Imaging* 36.1 (2017), pp. 86–97. DOI: 10.1109/TMI.2016.2593957 (cit. on pp. 58, 62, 69, 75, 76, 120, 128, 133).

[Twinanda 2019] Andru Putra Twinanda, Gaurav Yengera, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. "RSDNet: Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations". In: *IEEE Transactions on Med-*

ical Imaging 38.4 (2019), pp. 1069–1078. DOI: 10.1109/TMI.2018.2878055 (cit. on p. 62).

[Vant Hullenaar 2019] Cas DP Van't Hullenaar, Paula Bos, and Ivo AMJ Broeders. "Ergonomic assessment of the first assistant during robot-assisted surgery". In: *Journal of robotic surgery* 13 (2019), pp. 283–288. DOI: 10.1007/s11701-018-0851-0 (cit. on p. 49).

[Vercauteren 2007] Tom Vercauteren, Xavier Pennec, Ezio Malis, Aymeric Perchant, and Nicholas Ayache. "Insight into Efficient Image Registration Techniques and the Demons Algorithm". In: *Information Processing in Medical Imaging*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 495–506. ISBN: 978-3-540-73273-0. DOI: 10.1007/978-3-540-73273-0_41 (cit. on p. 87).

[Vondrick 2016] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. "Anticipating Visual Representations from Unlabeled Video". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 98–106. DOI: 10.1109/CVPR.2016.18 (cit. on p. 66).

[Voros 2007] Sandrine Voros, Jean-Alexandre Long, and Philippe Cinquin. "Automatic Detection of Instruments in Laparoscopic Images: A First Step Towards High-level Command of Robotic Endoscopic Holders". In: *The International Journal of Robotics Research* 26.11-12 (2007), pp. 1173–1190. DOI: 10.1177/0278364907083395 (cit. on p. 59).

[Wagner 2021] Martin Wagner, Andreas Bihlmaier, Hannes Götz Kenngott, Patrick Mietkowski, Paul Maria Scheikl, Sebastian Bodenstedt, Anja Schiepe-Tiska, Josephin Vetter, Felix Nickel, S Speidel, et al. "A learning robot for cognitive camera control in minimally invasive surgery". In: *Surgical Endoscopy* (2021), pp. 1–10. DOI: 10.1007/s00464-021-08509-8 (cit. on p. 70).

[Wagner 2023] Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David

M Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Annika Reinke, Carissa Reid, Tong Yu, Armine Vardazaryan, Chinedu Innocent Nwoye, Nicolas Padoy, Xinyang Liu, Eung-Joo Lee, Constantin Disch, Hans Meine, Tong Xia, Fucang Jia, Satoshi Kondo, Wolfgang Reiter, Yueming Jin, Yonghao Long, Meirui Jiang, Qi Dou, Pheng Ann Heng, Isabell Twick, Kadir Kirtac, Enes Hosgor, Jon Lindström Bolmgren, Michael Stenzel, Björn von Siemens, Long Zhao, Zhenxiao Ge, Haiming Sun, Di Xie, Mengqi Guo, Daochang Liu, Hannes G. Kenngott, Felix Nickel, Moritz von Frankenberg, Franziska Mathis-Ullrich, Annette Kopp-Schneider, Lena Maier-Hein, Stefanie Speidel, and Sebastian Bodenstedt. "Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark". In: *Medical Image Analysis* 86 (2023), p. 102770. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2023.102770. URL: https://www.sciencedirect.com/science/article/pii/S1361841523000312 (cit. on p. 133).

[Wang 2014] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. "Learning Fine-Grained Image Similarity with Deep Ranking". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1386–1393. DOI: 10.1109/CVPR.2014.180 (cit. on p. 66).

[Wang 2015] Xiaolong Wang and Abhinav Gupta. "Unsupervised Learning of Visual Representations Using Videos". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 2794–2802. DOI: 10.1109/ICCV.2015.320 (cit. on p. 66).

[Wang 2022] Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. "AutoLaparo: A New Dataset of Integrated Multi-tasks for Image-guided Surgical Automation in Laparo-

scopic Hysterectomy". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Cham: Springer Nature Switzerland, 2022, pp. 486–496. ISBN: 978-3-031-16449-1. DOI: 10.1007/978-3-031-16449-1_46 (cit. on pp. 75, 133, 146).

[Weede 2011] O. Weede, H. Mönnich, B. Müller, and H. Wörn. "An intelligent and autonomous endoscopic guidance system for minimally invasive surgery". In: *2011 IEEE International Conference on Robotics and Automation*. 2011, pp. 5762–5768. DOI: 10.1109/ICRA.2011.5980216 (cit. on p. 60).

[Wells 2019] Antonia C Wells, Magnus Kjellman, Simon JF Harper, Mikael Forsman, and M Susan Hallbeck. "Operating hurts: a study of EAES surgeons". In: *Surgical endoscopy* 33 (2019), pp. 933–940. DOI: 10.1007/s00464-018-6574-5 (cit. on p. 45).

[Wong 2023] Shing Wai Wong, Zhen Hao Ang, and Philip Crowe. "Improving ergonomics for the bedside assistant in robotic colorectal surgery". In: *Journal of Surgical Case Reports* 2023.1 (2023), rjad007. DOI: 10.1093/jscr/rjad007 (cit. on p. 49).

[Workum 2018] Frans van Workum, Laura Fransen, Misha DP Luyer, and Camiel Rosman. "Learning curves in minimally invasive esophagectomy". In: *World Journal of Gastroenterology* 24.44 (2018). PMID: 30510372, p. 4974. DOI: 10.3748/wjg.v24.i44.4974 (cit. on p. 49).

[Xie 2018] Annie Xie, Avi Singh, Sergey Levine, and Chelsea Finn. *Few-Shot Goal Inference for Visuomotor Learning and Planning*. 2018. arXiv: 1810.00482 [cs.LG] (cit. on p. 66).

[Xiong 2020] Ya Xiong, Yuanyue Ge, Lars Grimstad, and Pål J. From. "An autonomous strawberry-harvesting robot: Design, development, integration, and field evaluation". In: *Journal of Field Robotics* 37.2 (2020), pp. 202–224. DOI: 10.1002/rob.21889. URL: https:

//onlinelibrary.wiley.com/doi/abs/10.1002/rob.21889 (cit. on p. 78).

[Xu 2017] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. "End-to-End Learning of Driving Models from Large-Scale Video Datasets". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3530–3538. DOI: 10.1109/CVPR.2017.376 (cit. on p. 65).

[Xu 2022] Chi Xu, Baoru Huang, and Daniel S. Elson. "Self-Supervised Monocular Depth Estimation With 3-D Displacement Module for Laparoscopic Images". In: *IEEE Transactions on Medical Robotics and Bionics* 4.2 (2022), pp. 331–334. DOI: 10.1109/TMRB.2022.3170206 (cit. on p. 62).

[Yan 2020] Lei Yan, Wenfu Xu, Zhonghua Hu, and Bin Liang. "Multi-objective configuration optimization for coordinated capture of dual-arm space robot". In: *Acta Astronautica* 167 (2020), pp. 189–200. ISSN: 0094-5765. DOI: 10.1016/j.actaastro.2019.11.002. URL: https://www.sciencedirect.com/science/article/pii/S0094576519313797 (cit. on p. 78).

[Yang 2017] Guang-Zhong Yang, James Cambias, Kevin Cleary, Eric Daimler, James Drake, Pierre E. Dupont, Nobuhiko Hata, Peter Kazanzides, Sylvain Martel, Rajni V. Patel, Veronica J. Santos, and Russell H. Taylor. "Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy". In: *Science Robotics* 2.4 (2017), eaam8638. DOI: 10.1126/scirobotics.aam8638. URL: https://www.science.org/doi/abs/10.1126/scirobotics.aam8638 (cit. on pp. 36, 78).

[Yang 2019] Bohan Yang, Wei Chen, Zerui Wang, Yiang Lu, Jiayue Mao, Hesheng Wang, and Yun-Hui Liu. "Adaptive FOV Control of Laparoscopes With Programmable Composed Constraints". In: *IEEE Transactions on Medical Robotics and Bionics* 1.4 (2019), pp. 206–

217. DOI: 10.1109/TMRB.2019.2949881 (cit. on pp. 60, 102).

[Yasutomi 2023] André Yuji Yasutomi, Toshiaki Hatano, Kanta Hamasaki, Makoto Hattori, and Daisuke Matsuka. "Dual-Arm Construction Robot for Automatic Fixation of Structural Parts to Concrete Surfaces in Narrow Environments". In: *2023 IEEE/SICE International Symposium on System Integration (SII)* (2023), pp. 1–7. DOI: 10.1109/SII55687.2023.10039387 (cit. on p. 78).

[Ye 2017] Menglong Ye, Edward Johns, Ankur Handa, Lin Zhang, Philip Pratt, and Guang-Zhong Yang. *Self-Supervised Siamese Learning on Stereo Image Pairs for Depth Estimation in Robotic Surgery*. 2017. arXiv: 1705.08260 [cs.CV] (cit. on p. 76).

[Yengera 2018] Gaurav Yengera, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. *Less is More: Surgical Phase Recognition with Less Annotations through Self-Supervised Pre-training of CNN-LSTM Networks*. 2018. arXiv: 1805.08569 [cs.CV] (cit. on p. 62).

[Yu 2018] Tianhe Yu, Chelsea Finn, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. *One-Shot Imitation from Observing Humans via Domain-Adaptive Meta-Learning*. 2018. arXiv: 1802.01557 [cs.LG] (cit. on p. 65).

[Yu 2019] Tianhe Yu, Gleb Shevchuk, Dorsa Sadigh, and Chelsea Finn. *Unsupervised Visuomotor Control through Distributional Planning Networks*. 2019. arXiv: 1902.05542 [cs.RO] (cit. on p. 66).

[Yu 2020] Tong Yu, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. *Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition*. 2020. arXiv: 1812.00033 [cs.LG] (cit. on p. 62).

[Zach 2014] Christopher Zach. "Robust Bundle Adjustment Revisited". In: *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014, pp. 772–

787. ISBN: 978-3-319-10602-1. DOI: 10.1007/978-3-319-10602-1_50 (cit. on p. 90).

[Zach 2018] Christopher Zach and Guillaume Bourmaud. "Descending, Lifting or Smoothing: Secrets of Robust Cost Optimization". In: *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 2018, pp. 558–574. ISBN: 978-3-030-01258-8. DOI: 10.1007/978-3-030-01258-8_34 (cit. on p. 90).

[Zárate Rodriguez 2019] Jorge G Zárate Rodriguez, Ahmed M Zihni, Ikechukwu Ohu, Jaime A Cavallo, Shuddhadeb Ray, Sohyung Cho, and Michael M Awad. "Ergonomic analysis of laparoscopic and robotic surgical task performance at various experience levels". In: *Surgical endoscopy* 33 (2019), pp. 1938–1943. DOI: doi.org/10.1007/s00464-018-6478-4 (cit. on p. 45).

[Żelechowski 2023] Marek Żelechowski, Balázs Faludi, Murali Karnam, Nicolas Gerig, Georg Rauter, and Philippe C Cattin. "Automatic patient positioning based on robot rotational workspace for extended reality". In: *International Journal of Computer Assisted Radiology and Surgery* (2023), pp. 1–9. DOI: 10.1007/s11548-023-02967-2 (cit. on p. 49).

[Zhang 2020] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. "Content-Aware Unsupervised Deep Homography Estimation". In: *Computer Vision – ECCV 2020*. Cham: Springer International Publishing, 2020, pp. 653–669. ISBN: 978-3-030-58452-8. DOI: 10.1007/978-3-030-58452-8_38 (cit. on p. 119).

[Zhang 2023] Yu-Jin Zhang. "Camera Calibration". In: *3-D Computer Vision: Principles, Algorithms and Applications*. Singapore: Springer Nature Singapore, 2023, pp. 37–65. ISBN: 978-981-19-7580-6. DOI: 10.1007/978-981-19-7580-6_2. URL: https://doi.org/10.1007/978-981-19-7580-6_2 (cit. on p. 51).

[Zia 2021] Aneeq Zia, Kiran Bhattacharyya, Xi Liu, Ziheng Wang, Satoshi Kondo, Emanuele Colleoni, Beatrice van Amsterdam, Razeen Hussain, Raabid Hus-

sain, Lena Maier-Hein, Danail Stoyanov, Stefanie Speidel, and Anthony Jarc. *Surgical Visual Domain Adaptation: Results from the MICCAI 2020 SurgVisDom Challenge*. 2021. arXiv: 2102.13644 [cs.CV] (cit. on pp. 57, 58, 61, 70, 74, 76, 120).

[Zidane 2023] Iham F Zidane, Yasmin Khattab, Sohair Rezeka, and Mohamed El-Habrouk. "Robotics in laparoscopic surgery-A review". In: *Robotica* 41.1 (2023), pp. 126–173. DOI: 10.1017/S0263574722001175 (cit. on pp. 49, 50).

[Zuliang 2020] Feng Zuliang, Michael P. Feng, David P. Feng, and Carmen C. Solórzano. "Robotic-assisted adrenalectomy using da Vinci Xi vs. Si: are there differences?" In: *Journal of Robotic Surgery* 14 (2020), pp. 349–355. DOI: 10.1007/s11701-019-00995-2 (cit. on p. 49).