©Copyright 2025
Niveditha Kalavakonda Chandrasekar

Surgical Scene Understanding Towards Human-Centered Collaboration in Robotic Surgery

Niveditha Kalavakonda Chandrasekar

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Blake Hannaford, Chair

Maya Cakmak

Ryan Calo

Linda Shapiro

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

Surgical Scene Understanding Towards Human-Centered Collaboration in Robotic Surgery

Niveditha Kalavakonda Chandrasekar

Chair of the Supervisory Committee:
Blake Hannaford
Department of Electrical and Computer Engineering

Robot-assisted minimally invasive surgery combines the expertise of highly skilled surgeons with the enhanced precision and dexterity of assistive robots. Currently deployed surgical robots provide surgeons with enhanced visualization, filter out hand tremors, and incorporate data integration and analytics to facilitate improved patient care. Most surgical robots, however, are still constrained to being teleoperated by a surgeon, with limited advances towards integrating automation in the operating room. The use of automation in surgery has the potential to extend the current benefits of robotic surgery by integrating more consistent care, thereby reducing variability. While studies have explored automating subtasks performed by a surgeon, the role of the first assistant has often been underappreciated in the context of automation. The role of the bedside assistant is pivotal to ensuring the smooth progression of procedures, maintaining a clear surgical field, and providing critical support to the primary surgeon. To this end, delegating some subtasks performed by the first assistant to automated systems under cooperative autonomy can reduce the risk of human error and optimize resource allocation. Building on the viability of camera sensors as a reliable input, the goal of this thesis was to aid the development of an autonomous suction assistance tool by developing the infrastructure for surgical scene understanding. This work developed surgical segmentation models for performing efficient binary, parts, and instrument segmentation, established approaches for addressing the paucity of labeled data in surgical settings, and developed a framework to benchmark approaches to cooperative

autonomy through the lens of an assistive suction task.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables vi

Chapter 1: Introduction 1
1.1 Motivation 2
1.2 Research Outline and Contributions 5

Chapter 2: Locating Surgical Instruments in Robotic and Micro-neurosurgical Procedures 10
2.1 Background 10
2.2 Datasets 11
2.3 Robotic and Microneurosurgical Instrument Segmentation 13
2.4 Improving Instrument Type Identification Using Hybrid Transformer-CNN architecture 21

Chapter 3: Data Efficient Learning for Surgical Vision 28
3.1 Background 28
3.2 Active Learning 29
3.3 Balanced Expected Gradient Length with Density-Weighted Selection 35
3.4 Evaluation of Noisy Labels on Active Learning 38

Chapter 4: Towards generalizing performance for surgical scene understanding using self-supervised learning 46
4.1 Background 48
4.2 Method 52
4.3 Experiments 53
4.4 Analysis 54

Chapter 5: Human-Centric Surgical Robot Focused on Sub-task Automation . . . 55
5.1 Autonomy in Surgical Robotics 57

5.2 Metrics to Estimate Uncertainty and Communication of Competence Within Human-Robot Teams for Surgery . . . 59
5.3 A framework for benchmarking sub-task collaboration in surgical robotics . . 61
5.4 Conclusion . . . 74
5.5 Acknowledgements . . . 74

Chapter 6: Social Shaping of Surgical Robot Design . . . 76
6.1 Social Origins of Surgical Robotics . . . 77
6.2 Discussion . . . 87
6.3 Conclusion . . . 89

Chapter 7: Conclusion and Future Directions . . . 91
7.1 Future Directions . . . 92

Bibliography . . . 94

Appendix A: NeuroID Dataset Frames . . . 136

LIST OF FIGURES

Figure Number Page

1.1 The surgical scene captured during laparoscopic surgeries using an endoscope is starkly different from the scene during a microneurosurgical procedure recorded with a microscope due to lighting conditions, types of blur, types of instruments, etc. Most importantly, neurosurgical procedures are performed in a constricted operational volume. 3

1.2 With the introduction of assistive technology, surgical procedures have shifted from a team-centric interaction framework to a robot-centric framework. It is important to prioritize human-centered design to ensure effective team communication. (Illustrations adopted from pch.vector) 4

2.1 NeuroID Dataset consists of data that could accomplish two tasks - (1) binary instrument segmentation to distinguish instrument from background and (2) instrument identification where each instrument class has a separate label. The image shows (L-R) (i)the original frame, (ii)labels input using polygonal annotation in the LabelMe tool, (iii) Binary segmentation map and (iv)Instrument segmentation map 12

2.2 NeuroID has 8 different classes of instruments. The classes are imbalanced in the dataset, with suction being more prominent than other tools. 14

2.3 We visualize the EndoVis2017 and NeuroID datasets using t-SNE to demonstrate the similarities and differences in the data points between them. 15

2.4 Network architecture for segmentation. The encoder-decoder inspired network initially encodes a compact feature representation of the image and uses the corresponding upsampling layers to recover spatial resolution. 16

2.5 SurgSeg-HViT model architecture for a hybrid Transformer-CNN surgical instrument segmentation task 22

2.6 Dynamic Gating Block (DGB) 23

2.7 MFConvGating-SC Block enables multi-fusion at different resolutions and offers spatial and channel gating for improving model performance. 23

2.8 Qualitative results on MICCAI EndoVis2017 Dataset for Binary Segmentation 25

2.9 Qualitative results on MICCAI EndoVis2017 Dataset for Parts Segmentation 26

2.10 Qualitative results on MICCAI EndoVis2017 Dataset for Instrument Segmentation 27

3.1 Overview of Active Learning for Surgical Instrument Segmentation. A pre-trained machine learning model is used to query frames from an unlabeled data pool by using a selection criterion. 31
3.2 Qualitative results using Active Learning showing the uncertainty in the top batch of images selected for labeling by the model. 33
3.3 Active Learning Results: Overall performance of the model is achieved by using a maximum of 630 images (instead of 1500) for training a model with active learning. 34
3.4 Results of BEGL-DWS on EndoVis2017 and NeuroID datasets 38
3.5 We explore the downstream effects of a human oracle relying on AI-assisted annotation tools by simulating noisy labels 39
3.6 Noisy injection uses information from two techniques for noisy label identification – Noisy Completely at Random (NCAR) and Noisy at Random (NAR). NCAR simulates mislabeling where all classes are equally likely to be reassigned. NAR uses a weighted transition matrix for classes. 42
3.7 We evaluate the downstream effects of a human oracle relying on AI-assisted annotation tools by simulating noisy labels using NAR. The performance of a model trained on the entire OrganMNIST dataset is included in all the experiment results for comparison. 44

4.1 An overview of the self-supervised learning structure. Instead of relying only on annotated data, the use of pre-trained models and transfer learning enables improved generalization abilities. 47
4.2 Model Architecture for performing reference-based surgical instrument segmentation. 48

5.1 Experimental setup of the human-robot collaborative experiment setup. A user will *operate* a laparoscopic grasper instrument to perform peg transfer. This will be controlled by hand, and access to the taskbed will be via a port on a cardboard to simulate remote center of motion. (a) The RAVEN II surgical robot will work as the surgical assistant alongside the user to perform suction on a modified FLS task base. (b) Mimic Mantis will be used to teleoperate the robot. Teleoperation will be performed by a member of the study team. . 64
5.2 We design a modified peg-transfer task with cavities at the base of the board. This board is a tabletop experiment testbed to benchmark autonomy for surgeon and assistant tasks performed in parallel. 66
5.3 Overview of RAVEN II^{TM} Surgical System 68
5.4 We use a RAVEN tool to accommodate a neurosurgical suction tool by incorporating a 3D-printed adapter. Blood from cavities will be removed using a vacuum pump and filter attached to the suction instrument. 69

5.5 Location of Pegs on the task board, determined from point cloud data 70
5.6 Board Localization in camera frame 71

6.1 Teleoperated robotic surgery concept: (a) from Hugo Gernsback's illustration in 1925, (b) proposed by NASA Advanced Concepts and Missions Division in 1972 78

A.1 Sample frames from 5 videos in the NeuroID Dataset. These indicate the diversity in surgical conditions for the procedures recorded at the Harborview Medical Center - Seattle. 136

LIST OF TABLES

Table Number Page

2.1 Performance metrics for binary segmentation task. Best results are shown in **bold**, second best are underlined. 19

2.2 Performance metrics for instrument parts segmentation with per-class Dice scores (4 classes: background, shaft, wrist, claspers) 20

2.3 Performance metrics for multi-class instruments segmentation task (8 classes). Best results are shown in **bold**, second best are underlined. 20

2.4 Performance on Endovis2017 dataset for instrument type segmentation task . 24

3.1 Time and mIoU values for different active selection strategies. The estimated time taken to calculate uncertainty is indicated in seconds for 1 image. 34

4.1 Reference-based Surgical Instrument Segmentation on the EndoVis2017 Dataset. 54

5.1 Summary of autonomous surgical task and sub-task studies and the perceived level of autonomy upon successful *in vivo* implementation 58

ACKNOWLEDGMENTS

This thesis is the result of encouragement and kindness from many wonderful people I have been fortunate to learn from. Thanks to everyone who helped me get here; I could not have done this alone. First and foremost, I would like to express my sincere gratitude to my advisor, Prof. Blake Hannaford, for his mentorship, guidance, patience, and faith in me. He is the reason I moved to the University of Washington with the grand dream of pursuing research in surgical robotics. I am forever grateful for the opportunity to be part of the BioRobotics Lab (BRL). I would also like to express my gratitude to Prof. Asokan Thondiyath at the Indian Institute of Technology, Madras, for introducing me to the impactful field that is surgical robotics. My journey into graduate school would not have been feasible without his support and that of other robotics lab members in ED – they took me in when I knew very little about the field, and shaped my thinking about research and robotics in more ways than one. I would also like to thank the members of my doctoral thesis committee for their guidance, support, and patience: Prof. Maya Cakmak, Prof. Ryan Calo, Prof. Dieter Fox, and Prof. Linda Shapiro. I have been constantly inspired by each of them: Maya's passion for solving interdisciplinary problems that center humans has been my north star; Ryan's relentless advocacy for tech policy and his fostering of welcoming spaces for important conversations have consistently motivated me; Linda's exceptional academic standards in research, teaching, and service have set a high bar; Dieter has demonstrated through example how robotics research can extend beyond academic labs and have a significant impact in industry.

I extend my deep appreciation and heartfelt gratitude to our collaborators from the UW School of Medicine, Harborview Medical Center, and Seattle Children's Hospital. I am particularly thankful to Dr. Laligam Sekhar and his fellows for the opportunities to collaborate, observe surgeries, collect data, learn about how robots can positively shape

operating rooms, and so much more. At BRL, I have also been fortunate to be mentored by other PIs, including Prof. Howard Chizeck, Prof. Sam Burden, and Prof. Kim Ingraham. Their passion for research, teaching, and care for their students has had a profound impact on my career decisions. Howard, you were the first professor I assisted in teaching, and I learned a lot during all our interactions. Sam, it has been a whole journey under your guidance — from being your student to assisting in teaching and starting initiatives in the department. I learned to love control theory, math, and the value of a student-first approach to teaching because of your passion for it. Thanks for sharing your experiences in a way that makes people feel seen. Kim, you are amazing, and BRL is lucky to have you! I have learned to push through many a discomfort because of your encouragement, and I am grateful to have your support. Thank you all again for the invaluable support throughout my time here at UW; I am indebted to you folks.

I have been fortunate to be a part of a supportive lab that fosters an inclusive learning environment. My growth in graduate school comes from many conversations with present and past members of the BioRobotics Lab, including Andy, Yana, Haonan, Tin, Melody, Kyle, David, Astrini, Nava, Danying, Shan, Momona, Maneeshika, Amber, Andrew, Vidur, Harini, Zijie, Annika, Siena, Joey, Emmy, Jason, David, Mohammad, Mune, master's and undergraduate students, and many others. Thank you for creating an intellectually stimulating environment where I could learn about various aspects of robotics. I am especially grateful to Andy Lewis, Haonan Peng, and Melody Su for always being willing to help me out when I felt like I was hitting a wall. You helped me realize the importance of breaking problems down into sizable chunks by repeatedly questioning my core hypothesis. I am grateful for your guidance. While this journey had its set of challenges, what made it worthwhile was who I got to share it with. To my friends Alex, Brian, Itzue, Andre, Shashank, Shruti, Farah, Trang, Neeraja, Ekta, Shanti, Felix, Kedi, Jake, Matt, Mareldi, Nina, members of the Seattle Cattle and Electric Friz Grads teams, and all the others who I am going to regret not listing here in time– thank you for not letting me quit.

It also took me a long time to learn how to ask for help. There were many a faculty at

UW and external mentors who were empathetic and gave me the space to express myself. I thank Professors Eric Klavins, Payman Arabshahi, Eve Riskin, Anantram M. P., Josh Smith, Maryam Fazel, Amy Orsborn, Mo Li, Scott Hauck, and John Raiti at UW-ECE. I am also grateful for the support of Dr. Neil Tenenholtz and Sebastian Castro, who patiently answered all my questions about machine learning and robotics, no matter how big or small. The staff at the UW-ECE department made this place welcoming from day one and have been instrumental in ensuring my well-being throughout my time here. Many, many thanks to Stephanie Swanson, Jennifer Huberman, Mack Carter, May Lim, Brenda Larson, Jessie Allen-Navarre, Ryan Hoover, Christie Peralta, Travis Salig, Wayne Gillam, Ary Prasetyowati, Rodney Wells, Aleesha Wiest, Chris Overly, Cole Collins, Erika Toto, Lauren Van Fossen, Alma Royeca, and others. A special shoutout to Stephanie, Jennifer, and Mack is long due; thanks for saving my sanity through many uncertainties during COVID and beyond. Thanks for everything! Also, May, the number of lives you've positively shaped is one too many. I am grateful to be on that list.

I discovered my passion for teaching during my time at UW. I have had the opportunity to teach EE543 Models of Robot Manipulation twice (Spring 2024 and Winter 2025). Thank you to all my students who were patient with me while I learned the ropes of the job, and provided constructive feedback that helped me realize how much I enjoyed robotics education. I have benefited from the generosity of many mentors during my search for the next step in my career. I am especially grateful to Dr. Taylor Kessler-Faulkner, Prof. Michelle Hickner, Prof. Rania Hussein, Prof. Tamara Bonaci, and Prof. Sep Makhsous, who helped me communicate my passion for teaching to faculty search committees during an unusual year for academia

Outside of UW, I have learned a great deal from my internship and co-op teams. I owe a debt of gratitude to Todd Hester, Sheng Chen, Krishna Sridhar, Josef Spjut, Nick Jong, Ben Boudaoud, Karla Vega, Joanna Arreaza-Taylor, Ben Shapiro, Jim Harmon, and other members of my teams. These experiences shaped the direction of my research at the university and made it that much better. I have been inspired by the wonderful researchers

I have been fortunate to learn from and turn to for support in the Women in Computer Vision (WiCV) and Machine Learning for the Developing World (ML4D) communities, as well as the RSS Pioneers 2021 Cohort and the EECS Rising Stars 2023 Cohort. Additionally, my co-organizers and speakers from Talking Robotics have had an immense impact on my understanding of the robotics research landscape beyond surgical robotics; for that, I am grateful. I have also had the opportunity to learn from some phenomenal women who are passionate about community, machine learning, and ethics. I would like to extend my thanks to Amanda Casari, Kathy Pham, Sheila Oh, Sarah Hubbard, Margaret Maynard-Reid, Shivani Patel, Mengyuan Liu, Arushi Prakash, Kristin Mussar, Ivana Milovanovic, and everyone else from the WiDS-PS and GDG organizing teams.

Finally, I want to thank my friends and family back in India (and other parts of the world), who have been a constant source of strength and joy in my life. They believed in me, even during times when I doubted myself. Without your encouragement and continuous support, this journey would have been impossible for me. Thank you for your unwavering love and support.

DEDICATION

to my family

Chapter 1

INTRODUCTION

Over the last decade, advancements in the field of machine learning have contributed to the improved performance of robots in everyday environments [1] [2] . Autonomous robots, in contrast to manipulators currently used in manufacturing, are tasked with solving a variety of problems under uncertainty and incomplete knowledge [3]. This multifaceted role encompasses the process of (a) perceiving the environment in which they are placed, (b) planning that is predicated on processing the perceived input, and (c) executing actions at each time step. Current robotic systems are capable of discerning their environment using various sensors, the most ubiquitous being vision [4]. Robotic vision uses input from image sensors to perform a range of tasks such as object detection, semantic segmentation, depth estimation, and 3D reconstruction, among others [5–7]. Output from these perception tasks is used for decision-making to drive robot action.

Paralleling the advancements of machine learning, the integration of digital technology for medical applications has permeated different stages of patient care [8]. The utility of computer-assistance in procedural fields (such as surgery and endoscopy) and for aftercare has been consistently evolving [9]. This has been observed through the use of technology for monitoring patient vitals [8], novel imaging techniques [10], and even robotic technology [11]. The use of computer-assisted interventions in the operating room has helped with early diagnosis of diseases [12], and increased efficiency [13] and advances in surgical technique [14]. The field of surgical robotics has seen rapid progress since the mid-1980s, translating initial efforts in lab studies to widely marketed, commercially successful devices [14]. Robotic surgery offers improved ergonomics [15,16] for the surgeon, more degrees of freedom compared to minimally invasive laparoscopic surgery, decreases surgeon's hand tremor that could result in undesired interactions with delicate tissue or organs [17], and facilitates enhanced visualization capabilities [18]. Surgeons are more willing to integrate

robotic surgery into the operating room, and there has been a steady growth in its adoption by hospitals in recent years [19].

In this introductory chapter, we provide the rationale and significance of this thesis, along with an overview of the subsequent chapters. First, we motivate surgical scene understanding, a process that involves perceiving and analyzing dynamic interactions observed during a surgical procedure, from a clinical perspective. Next, we highlight the primary questions that are proposed and how they have been addressed in this thesis, referencing the chapters that are relevant to each question. We conclude with information about the resulting contributions, potential extensions, and future directions for this work.

1.1 Motivation

Despite the potential of surgical robotics, they have disproportionately advanced some areas of surgery more than others. The fields of general surgery, urology and gynecology have benefited the most from Robot-assisted Minimally Invasive Surgery(RMIS). The da Vinci Surgical System [20,21], a laparoscopic robot by Intuitive Surgical Inc., has seen a consistent increase in its adoption, for instance, observing an 18% increase in its use in 2022 compared to 2021 [22]. The fields of thoracic surgery, cardiac surgery, orthopedics, and retinal surgery are some others that have seen increased adoption of surgical robots [23]. While the exploration of advanced techniques in surgical robotics is further advanced in laparoscopic, orthopaedic, and ophthalmic surgeries, fields like neurosurgery are relatively less explored due to the constrained operative space. An analysis of innovations in endovascular and cerebrovascular neurosurgery detailed in [24] shows that most advances have been related to imaging modalities and surgeon-controlled navigation systems such as the Pathfinder, Neuromate, and SpineAssist. Similarly, despite the methodological advances being made in surgical scene understanding, the surgical specialties addressed in datasets are rather limited [25].

In our work, we study surgical scene understanding as a step toward deploying assistive robotics for neurosurgery in Chapter 2. The integration of digital technology into the operating room has resulted in the collection of a large volume of surgical data. There is untapped potential in using the collected data towards improved generalizability for Robot-

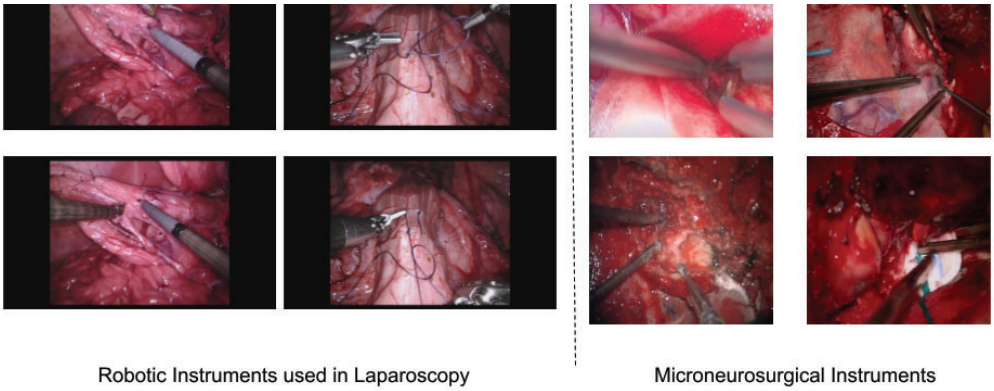


Figure 1.1: The surgical scene captured during laparoscopic surgeries using an endoscope is starkly different from the scene during a microneurosurgical procedure recorded with a microscope due to lighting conditions, types of blur, types of instruments, etc. Most importantly, neurosurgical procedures are performed in a constricted operational volume.

assisted surgery and, more broadly, Computer-Assisted Surgery. To this end, we explore data-efficient learning techniques for surgical scene understanding in the context of current laparoscopic and microneurosurgical procedures in Chapter 3 and 4.

Additionally, the types of surgeon-robot interactions studied have either focused on teleoperation or supervised control [26], where the surgeon is in control of robot's actions, or traded control [27, 28] or fully autonomous control [29], where a robot is responsible for a surgical task/sub-task. The viability of simultaneous collaborative control between a surgeon and robot to complete a shared goal has been relatively unexplored [30]. The work done through this thesis is focused on applications in human-robot interaction experiments that center the surgeon's needs while the robot assists to complete a shared goal by working within the same surgical workspace.

An important component of successful robot deployment is the seamless interaction between humans and robots. To advance cooperative assistive automation for robotic surgery, we need to develop components that learn from surgical interactions, create improved imaging- and instrument-oriented subsystems, and establish a surgeon-centric interface to integrate validated computational insights. Postulating that more complex proce-

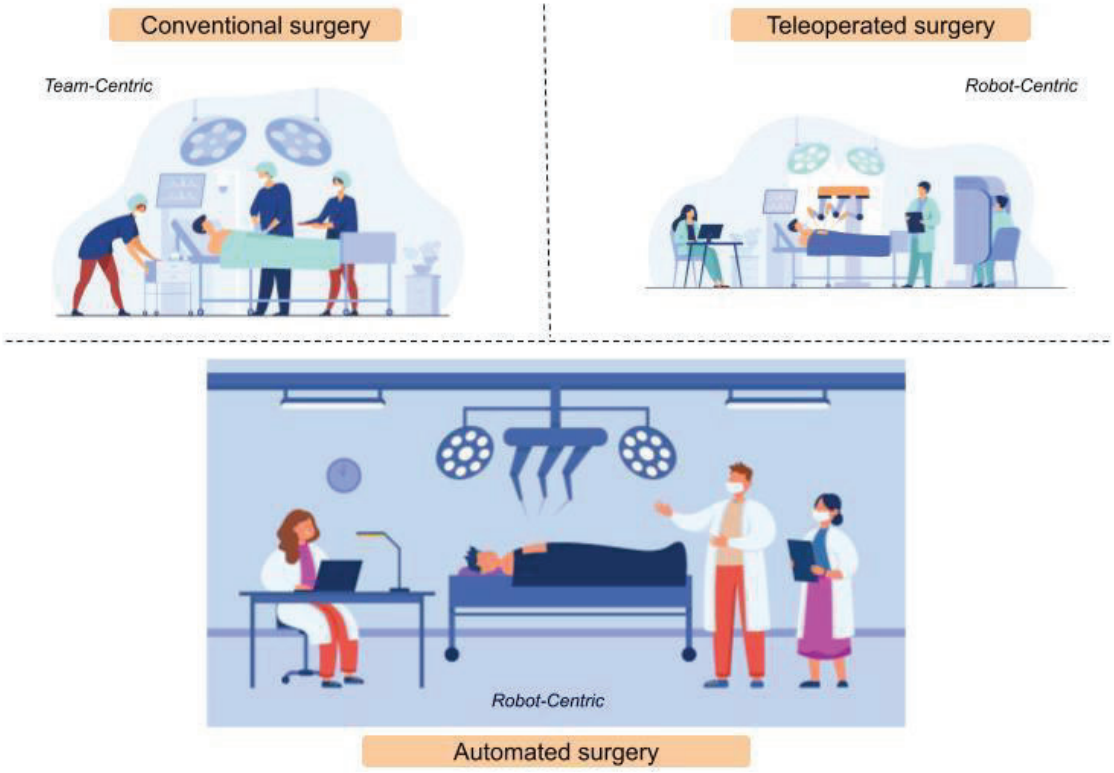


Figure 1.2: With the introduction of assistive technology, surgical procedures have shifted from a team-centric interaction framework to a robot-centric framework. It is important to prioritize human-centered design to ensure effective team communication. (Illustrations adopted from pch.vector)

dures would benefit most from incorporating supervised automation and robotics through improved precision, minimal invasion, and consistent surgical technique, it is crucial to design systems that are not only precise and safe but also centered on user needs. Furthermore, with the increase in the quality of care and technique, it is essential to provide interventions with reliable outcomes and minimal variability between the teams providing them. Collaborative automation has the potential to be the next generation of assistive tools for surgeons, focusing on the automation of surgical subtasks and centering on user needs. To achieve this goal, it is important to comprehend the spatial relationships between the camera viewpoint (endoscope or microscope), surgical instruments, and patient anatomy. The main contribution of this thesis is to provide models that can generate reliable surgical scene understanding (techniques developed in Chapters 2, 3, and 4) to move closer to assisting with autonomous suction using a surgical robot while working collaboratively with a surgeon in a shared workspace. This system is described in Chapter 5.

1.2 Research Outline and Contributions

I aim to design systems and sub-modules that facilitate autonomous collaborative robotic assistance in surgery with a specific focus on cooperative task completion. Towards this goal, I use tools from computer vision and machine learning to address three questions – (1) How do we identify the objects present in a safety-critical scene such as surgery? (2) How do we achieve generalizable performance even when there is a scarcity of labeled surgical data? (3) How do we establish collaboration between surgeons and robots for cooperative autonomy?

1.2.1 Phase 1: Evaluate the Potential of Image Sensors for Identifying Surgical Instruments in Robotic Procedures

Most surgical procedures rely on camera technologies such as endoscopes or microscopes to relay internal anatomy to the surgeon and surgical team when there is no direct visual line of sight. The images and videos recorded by these cameras serve as a rich source of information from surgical sites, including information about the surgical process, patient anatomy and

the types of tool-tissue interactions that occur during the procedure. Extracting information from this data-rich input using machine learning can help advance computer-assisted and robot-assisted interventions. Automated instrument recognition in surgical frames can aid the development of intelligent applications in instrument tracking, pose estimation, augmented reality overlay, visual servoing, and analysis of surgical phase [31]. In addition to improving the performance of downstream surgical tasks, the position information of instruments also helps analyze intraoperative data, identify potential collisions, and provides context-aware guidance for surgeons by reasoning about surgical phases [11], thus potentially reducing human errors and enhancing safety. Tracking instruments also facilitates addressing myriad challenges in computer-assisted interventions. For example, the tracking of surgical instruments can be used for interventional workflow analysis and overlaying additional graphical information.

- **Semantic Segmentation of Surgical Scenes** Identifying instruments in an image using bounding boxes is not sufficient due to the coarse boundaries generated around the regions of interest. The dense prediction of tool versus background through semantic segmentation [32] is crucial in surgical contexts as it enhances safety by parsing the whole scene without suppressing occluded or slanted objects. Semantic Segmentation is relatively better positioned to avoid false positives, generates more precise image boundaries, is more adaptable to varying conditions within the surgical site, and facilitates further processing from the extracted information. Therefore, in phase 1, we focused on generating pixel-wise labels to delineate surgical instruments from their background and from each other using supervised machine learning. Deep Learning (DL) models, a subset within machine learning, learn end-to-end mapping from input to pixel-level labels and avoid the need for manual feature engineering, a limitation of classical computer vision. We demonstrate that a hybrid transformer and convolutional neural network (CNN) approach can provide efficient and robust segmentation for surgical instruments.

1.2.2 Phase 2: Demonstrate the Viability of Using Machine Learning under a Paucity of Labeled Data

Supervised machine learning techniques require a large volume of labeled data to train models. While unlabeled video data from surgical procedures is relatively easy to obtain, it is a challenge to acquire data labels for surgical tasks due to the domain-knowledge required from annotators. This limits the applicability of supervised deep learning techniques for different types of surgical scenes. Tasks such as semantic segmentation are also labor intensive due to the fine-grained annotation required for training. Pixel-wise annotations are expensive and time-consuming to generate, especially factoring in time from highly qualified professionals. Is surgical instrument recognition possible in domains where manual annotations are scarce? We address this through two different approaches as described in Aim 2a and Aim 2b.

- **Aim 2a: Interactively Querying Data Labels Using Active Learning**

  A limitation to obtaining labeled data that is generalizable is the imbalanced nature of classes that occur in these datasets and the redundancy in available information from sequential frames [33]. This introduces a bias in performance towards more commonly represented classes such as background or frequently used instrument types such as suction. Therefore, selecting frames that would be most informative for our machine learning framework, while reducing the amount of training data required, would be beneficial. Towards this goal, we develop an approach using active learning, a machine learning technique in which the learning algorithm attempts to maximize the performance gain for a model while querying the human-annotator for the fewest data labels [34]. Does interactively querying data for annotation result in improved performance of a machine learning model with less labeled data? This question is addressed in Chapter 3.

- **Aim 2b: Improving Generalizability of Downstream Tasks Using Foundation Models and reference-based surgical instrument segmentation**

Effective representation of surgical image data has the potential to benefit a number of downstream tasks such as surgical phase recognition and objective skill analysis. The scale of labeled data available to learn from using supervised training is limited in the medical field. A pre-trained model, trained on the large unlabeled dataset, allows us to freeze the initial layers and finetune on a smaller labeled dataset for our task, benefiting from transfer learning [35]. In Chapter 4, we demonstrate the use of foundation models to perform fine-grained referring surgical instrument segmentation.

1.2.3 Phase 3: Develop a framework for Human-Centric Surgical Robot Focused on Subtask Automation

Surgical robots can offer different levels of automation [36] and hence various modes of human-robot interaction [37]. The classification schemes for automation proposed by the research community [36], however, have been centered at a task-level and fall short on accounting for surgeon needs [23]. In the operating room, surgeons work with a surgical team to effectively execute a surgical procedure. The primary surgeon controls the surgical robot from a remote console while a bedside assistant provides support by giving and grabbing suturing needles, extracting specimens, clipping when needed, and providing suction and irrigation to maintain visibility by clearing blood from the surgical field. This relationship is a key feature of successful surgical procedures.

- **Aim 3a: Framework for benchmarking Cooperative Autonomy in Surgery** Automation of robotic assistants for surgery prioritize shared control, where the robot alternates with a primary surgeon to perform the task, under the guidance of the surgeon [27]. Other research in surgical robotics is moving towards supervised autonomy by automating surgical tasks performed by surgeons through learning from demonstration [38]. These methods currently do not account for the instruments used by a first assistant operating simultaneously. There are potential benefits to autonomously assisting the surgeon to safely perform surgery and to supporting interactions between a surgeon and the robot [23]. These include automation of repetitive tasks and reducing fatigue for the surgical team.

Effective implementation of a collaborative robotic assistant will need three important attributes: surgical scene understanding, plan generation, and execution of surgical action. This project focuses on the essential metrics to develop surgeon-centric assistive robots and describes a framework to benchmark tasks where surgeons and robots perform actions in parallel towards a shared goal. Our system is described in Chapter 5.

- **Aim 3b: Social Shaping of Surgical Robotics and the Implications on Surgical Robot Design**

  Surgical robots have revolutionized the way surgery is performed. However, they have also displaced the primary surgeon from near the patient cart to a surgical console, resulting in a shift in the surgeon's direct control at the bedside and emphasizing the need for a proficient surgical team [39]. As robot-assisted surgery is used more regularly in surgical care, there has also been an increase in the desire to integrate surgical automation. Recognizing users as continually evolving in the construction of technology is vital while accounting for ways in which technology plays a role in the construction of team dynamics. This research views the robotic operating room as a socio-technical system where the surgical team adapts to technology by modifying their communication. How does the surgical robot design impact team communication in the operating room? What are factors that need to be accounted for to center needs of humans in surgical robot design? We attempt to answer these questions through Aim 3b in Chapter 6. We present an analysis of literature on surgical robot design to establish the factors relevant to automated assistance and teamwork.

Each chapter contains a focused literature review. We conclude by providing a conclusion and future directions in Chapter 7.

Chapter 2

LOCATING SURGICAL INSTRUMENTS IN ROBOTIC AND MICRO-NEUROSURGICAL PROCEDURES

Digital cameras have revolutionized the way surgery is performed today, where they are ubiquitous in the modern operating room. Endoscopes and labor scopes are commonly used in robotic surgery and microscopes are used in microsurgery. Additionally, a variety of cameras are used for both therapeutic and diagnostic procedures [31]. We are able to collect intraoperative information from the operating room with the use of cameras, but this data has been under utilized outside the scope of educational references.

Computer vision applied to surgery has the potential to facilitate safe surgical interventions through pre-operative planning, surgical navigation, assistance and guidance during surgical procedures, and for post-operative analysis [40]. Surgeons rely on endoscopic images to operate on patients during minimally invasive surgeries (MIS) due to direct access to the surgical site being limited. In the last decade, robotic technologies have been used to overcome the limitations of conventional MIS tools and provide increased dexterity, control, and maneuverability required for precise surgical tasks [14].

2.1 Background

Depending on the type of procedure and the stage within a surgery, there could be multiple instruments present in the field of view to operate on a surgical site. These instruments are constantly in motion, tending to both planned and unexpected events that occur over the course of a procedure. A challenging aspect of performing surgical scene analysis using images are the presence of confounding factors in surgical settings that include motion blur due to rapid movement, bleeding, specular reflections, occlusions, and smoke [41]. Computer-assisted surgery has been increasing in popularity due to the myriad benefits it provides surgeons. Given the highly dynamic nature of operating rooms, providing redundant information could enhance safety while providing a flexible working environment for surgeons.

Imaging information from surgical procedures provide information about the events at different temporal levels. This could be classification of procedures at the video-level [42], for recognizing surgical phases at a coarse level [43], to recognize the steps within a phase [44], for event recognition (such as bleeding) [45] or content recognition at a frame-level or inter-frame level using semantic segmentation or tracking [46, 47].

Surgical tool segmentation and identification helps alleviate some of these challenges by locating the position of instruments in each frame. Tool tracking is an enabling technology and serves as a core component to realizing autonomous assistance in surgery from cognitive or physical support (level 1) to high autonomy (level 4) or even fully autonomous surgery (level 5) [23]. These tasks could potentially provide real-time visual feedback to surgeons [48], especially in robotic surgery, by highlighting the position and movement of surgical instruments within an operative field and enabling spatial situational awareness. They could also enhance surgical precision, reduce human error and optimize the surgical workflow. Automated segmentation and tracking of surgical instruments during surgery, however, can be a challenging task due to the dense placement of objects in the operating room, making it harder to introduce tracking devices, and/or the variation in lighting, shadows, instrument overlap, reflections from tissue on instrument (and vice versa), a range of textures, and occlusions from blood, fogging or surgeon interference in the frame making it challenging for vision-based techniques [49].

2.2 Datasets

The volume of data generated in an operating room is large but not all information acquired is recorded or permanently stored for learning purposes [50]. Despite stereoscopic endoscopes and microscopes providing high resolution streams at 30 Hz, efficient storage of high-resolution video data for surgical procedures is challenging. The data collected during procedures can be harnessed for research purposes, leading to the development of improved surgical techniques, training programs, and best practices. This, in turn, fosters continuous improvement in surgical practices and instrument designs. The untapped potential of surgical data in the operating room warrants further exploration into its capabilities for different types of surgical applications.

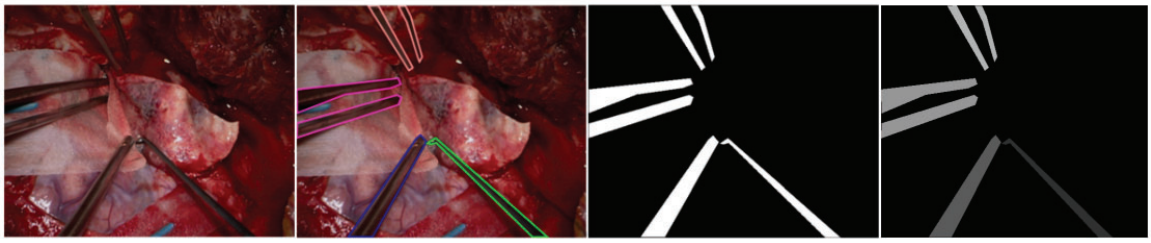


Figure 2.1: NeuroID Dataset consists of data that could accomplish two tasks - (1) binary instrument segmentation to distinguish instrument from background and (2) instrument identification where each instrument class has a separate label. The image shows (L-R) (i)the original frame, (ii)labels input using polygonal annotation in the LabelMe tool, (iii) Binary segmentation map and (iv)Instrument segmentation map

NeuroID dataset

Neurosurgical tumor resection involves a delicate balance of excising maximal amount of tumor while also ensuring there is no risk of neurological damage. Surgical procedures for neurosurgery use instruments that enable increased precision. Therefore, the type of instruments used in microneurosurgical procedures are starkly different from the more commonly explored laparoscopic surgical instruments which constitue a majority of surgical instrument datasets [33, 51].

We generated the NeuroID dataset [52] from six videos recorded in-vivo at the Harborview Medical Center in Downtown Seattle. An application to the Institutional Review Board at the University of Washington was approved to collect de-identified data with patient consent. The surgical procedures involved left frontal cavernoma, right sphenoid wing meningioma, left petroclival chondrosarcoma, brainstem cavernoma, and right sphenoorbital meningioma. These procedures used up to five instruments simultaneously in the surgical field and the videos included segments where there were no surgical instruments in the frame.

Each video has a resolution of 720 x 480 px and runs at 29.97 frames per second. The images for the surgical tools dataset were collected every 14^{th} frame to record information

at an approximate rate of 2 frames per second. A total of 3000 images have been manually annotated with the help of neurosurgery fellows. Bounding polygons were used to generate segmentation ground truth for tool versus tissue and labels for each tool-type. The dataset contains 8 different classes of instruments labeled for purposes of instance segmentation. The instrument categories include electrocautery, pickup, grasper, cutter, curette, drill, suction, and others. We generated annotations for the dataset using the LabelMe annotation tool [53]. A sample annotation is shown in Fig. 2.1 and a distribution of instrument precense in a representative subgroup is depicted in Fig. 2.2. We also have an expanded NeuroID-v2.0 dataset, which includes frames from additional surgical procedures that are unlabeled. This results in a dataset of 6000 images in total.

Challenges: On the data collection front, obtaining data for neurosurgical procedures was difficult due to personnel availability, mechanics of obtaining patient consent, and fewer surgeries actively involving tools simultaneously operated by the surgeon and the surgical assistant. Surgeries with instruments solely operated by the surgeon were not as rich in information and were less relevant to our goal of collaborative surgical assistance. Additionally, due to idiosyncratic depth of tumor (between 1 to 5cm) and its location in the frame, the level of focus blur and extent of blurred area are also highly variable among the videos.

2.3 Robotic and Microneurosurgical Instrument Segmentation

2.3.1 Prior Work

Broadly, there are two categories of approaches for instrument detection: marker-based versus marker-less. Marker-based approaches make use of physical modifications or indicators on an instrument to identify the type of instrument and its location. Using marker-based approaches would add to additional costs for manufacturing and sterilization of markers instruments. It also adds to the complexity of a procedure by requiring constant calibration for trackers to obtain position information relative to the global coordinate frame. Scalability of the solution is hence more applicable to marker-less solutions [54]. Bouget et al. [55] detect surgical tools by learning the local appearance and global shape from the training

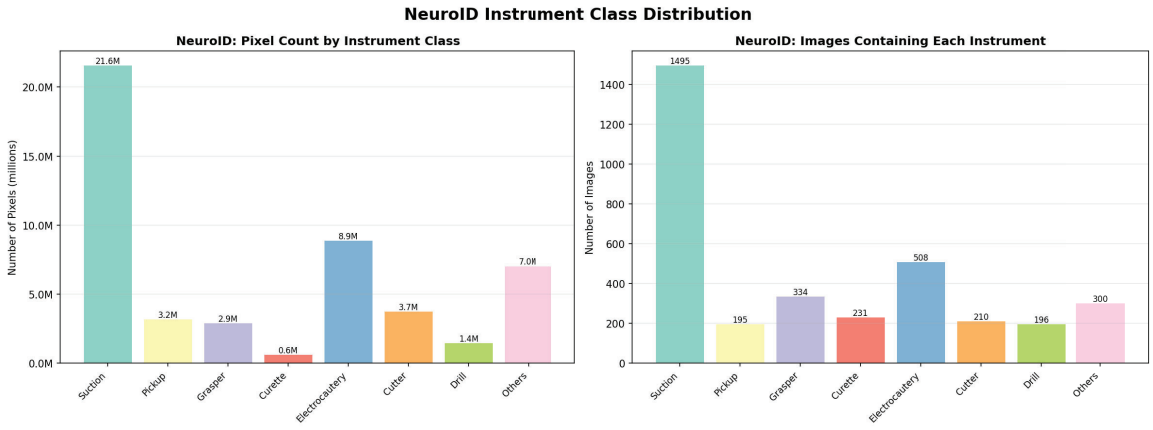
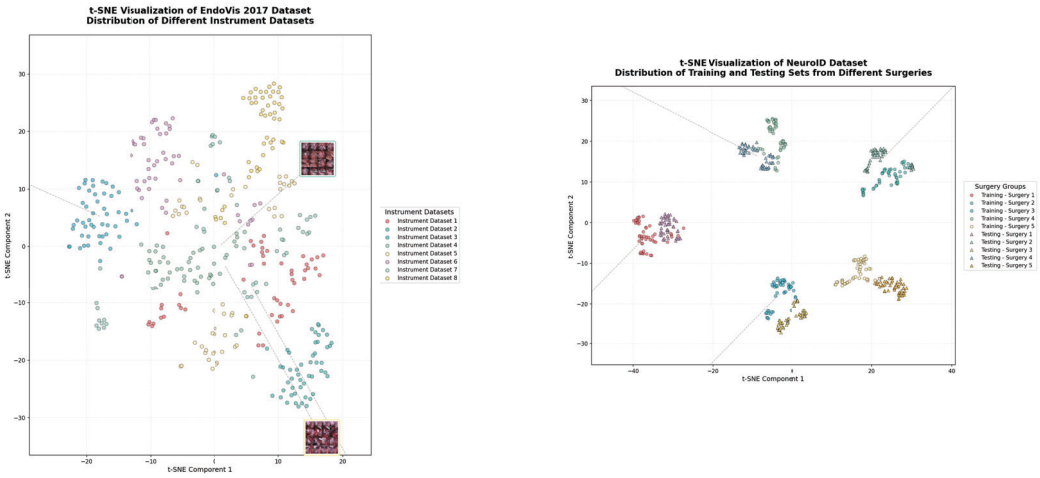


Figure 2.2: NeuroID has 8 different classes of instruments. The classes are imbalanced in the dataset, with suction being more prominent than other tools.

data. A random forest model is trained over ten feature channels and the shape model is learned using linear Support Vector Machine. The parameters are searched exhaustively and hence slow for run time implementation. This work is also one of two cases that developed a neurosurgical dataset for identifying instruments at the time of publication. There has been one other publication following ours that has worked on neurosurgical instruments [56]. A list of surgical tool datasets currently available publicly are summarized in [57].

Recent research has been more oriented towards using machine learning and deep learning for surgical tool segmentation. [58] proposed using classical computer vision techniques where the CIELAB color space and Otsu's thresholding [59] were used to identify instruments and hemorrhage candidate areas in an image. In recent years, a number of image-based techniques to identify surgical tools have been proposed, primarily using convolutional neural networks [33, 60–63]. [64] uses Convolutional Neural Networks with an encoder-decoder architecture. Pakhomov et al. [65] modified a deep residual learning network ResNet-101 and used dilated convolutions with stride to perform a binary segmentation task for differentiating surgical instruments and tissue. The output image dimensions are restored by incorporating a deconvolutional layer. Both these methods were trained on datasets with pixel-wise segmentation, restricted to the field of Minimally Invasive Surgery.



(a) EndoVis2017 t-SNE

(b) NeuroID t-SNE

Figure 2.3: We visualize the EndoVis2017 and NeuroID datasets using t-SNE to demonstrate the similarities and differences in the data points between them.

The other challenge for using deep learning on current datasets is the large number of parameters in the network used in proportion with the smaller size of the dataset, which may lead to overfitting. In this chapter, we present our work in developing two efficient neural network approaches to surgical instrument segmentation. The first model is based only on convolutional neural networks (CNN), which are great at performing on local features, while the second approach called SurgSeg-HViT is a hybrid Transformer-CNN model that combines the advances in vision transformers [66] with CNNs.

2.3.2 Dataset

For the rest of this thesis, we evaluate our methods on a more broadly available benchmark dataset from the Endoscopic Vision Instrument Segmentation Challenge [67] at the Medical Image Computing and Computer Assisted Invervention (MICCAI) 2017 conference to ensure standardized comparisons with the state-of-the-art methods. This dataset will be referred to as EndoVis2017 in the subsequent sections.

A key difference Endovis2017 and NeuroID is in the diversity of data points included in

the datasets. This is captured in a t-SNE plot of the two datasets in Fig. 2.3. The data points from the different videos for EndoVis2017 lie close to each other, whereas the datapoints for endovis are farther apart. This also indicates the practicality of using NeuroID to evaluate out-of-distribution scenarios by holding out specific videos.

2.3.3 Methods

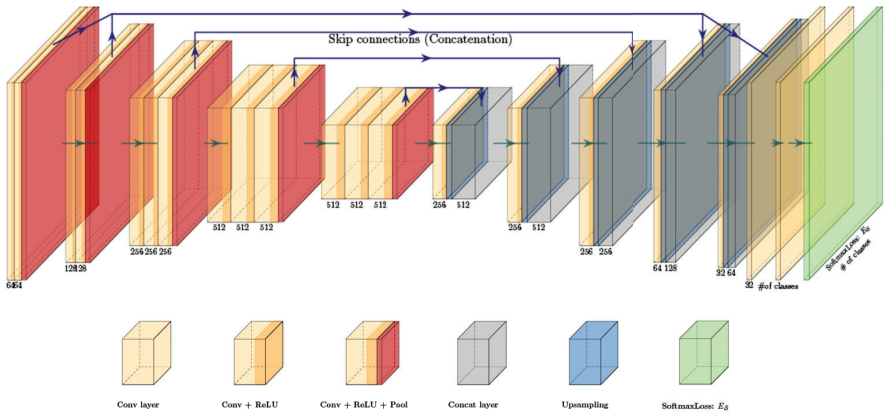


Figure 2.4: Network architecture for segmentation. The encoder-decoder inspired network initially encodes a compact feature representation of the image and uses the corresponding upsampling layers to recover spatial resolution.

In this work, we used models following an encoder-decoder structure as proposed in U-net [68] for the tool vs non-tool binary segmentation task and multi-class segmentation tasks such as parts and instrument type identification.

The encoder network consists of successive convolutional layers, pooling layers and Rectified Linear Unit (ReLU) activations, capturing a compact feature map in an encoded latent

representation. The pooling layers from the encoder are replaced with upsampling layers in the symmetric decoder network for recovering spatial information. The concatenation of higher resolution features from the downsampled path with the features in the upsampled section (Fig. 2.4) provides precise localization. The performance of the default encoder and decoder were rather limited. By leveraging the benefits of transfer learning [69] we evaluated variations in U-net, including a VGG11 encoder, VGG16 encoder network [70], and a ResNet34 encoder to be similar to LinkNet34 [71]. Since these networks were still quite large, we proposed using a lightweight backbone by incorporating a more versatile MobileNetV2 [72] backbone. The U-net with VGG16 backbone consists of 3x3 kernels and uses 1x1 convolutions intermittently to change the dimensionality in the filter space whereas the MobileNetV2 network presents an inverted residual structure.

All the networks were pre-trained on the ImageNet dataset [73]. For upsampling in the decoder network, we used bilinear interpolation for the Vanilla U-net, fractionally strided convolutions/transposed convolution with the for VGG networks as proposed by Ternaus-Net [74], and using nearest neighbor interpolation for Mobile-UNet.

MobileNetv2 uses depthwise separable convolutions, linear bottleneck between layers and shortcut connections between the bottlenecks. It uses fewer parameters and runs faster in comparison to VGG-16, without a large loss in accuracy. The obtained pre-trained weights were initially trained for faster inference speeds as opposed to higher accuracy. This initially resulted in relatively lower performance. We updated our decoder for MobileUNet based on results from [75] to create a computationally efficient solution. This uses a data-dependent upsampling technique called DUpsampling that incorporates better feature aggregation and downsamples the fused features to the lowest resolution before merging them.

The energy function is minimized to maximize probability of accurate predictions. The loss function is binary cross entropy loss for binary segmentation and a combination of dice and categorical cross entropy loss for parts and instrument type segmentation. We perform K-fold cross-validation to avoid overfitting. The dataset was augmented [76] using horizontal flip, vertical flip, normalize, padding and random crop to increase its size and learn invariance properties. A threshold of 0.3 was used to determine if the pixel was tool or background. All pixel values below the specified threshold were set to 0, while all

values above the threshold were set to 255 to produce final prediction mask for the binary segmentation task, similar to [61]. The networks were trained for 50 epochs, with a batch size of 4. The networks used Adam optimizer with an *alpha* of 0.0001.

2.3.4 Analysis

On applying the instrument segmentation algorithm detailed above, it was possible to evaluate the performance of the model on all three tasks.

To quantify the relative performance of the neural networks with respect to the hand-crafted heuristic baseline [67], the Dice coefficient and Intersection over Union (IoU) metrics were used to calculate quality of binary segmentation (Table 2.1) and multi-class segmentation for parts in Table 2.2 and instrument type in Table 2.4.

The Intersection over Union (IoU) for a single class c is defined as:

$$\text{IoU}_c = \frac{|P_c \cap G_c|}{|P_c \cup G_c|} = \frac{TP_c}{TP_c + FP_c + FN_c} \tag{2.1}$$

where P_c and G_c are the predicted and ground truth regions for class c, respectively. TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives for class c.

For pixel-wise computation:

$$\text{IoU}_c = \frac{\sum_{i=1}^{N} p_{i,c} \cdot g_{i,c}}{\sum_{i=1}^{N} \max(p_{i,c}, g_{i,c})} \tag{2.2}$$

where $p_{i,c}$ and $g_{i,c}$ are binary indicators (0 or 1) for pixel i belonging to class c in prediction and ground truth, respectively, and N is the total number of pixels.

The mean IoU across all C classes is:

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \text{IoU}_c \tag{2.3}$$

Similarly, the Dice coefficient (also known as F1 score or Dice-Sorensen (DSC) coefficient) for a single class c is:

$$\text{Dice}_c = \frac{2|P_c \cap G_c|}{|P_c| + |G_c|} = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \tag{2.4}$$

For pixel-wise computation:

$$\text{Dice}_c = \frac{2\sum_{i=1}^{N} p_{i,c} \cdot g_{i,c}}{\sum_{i=1}^{N} p_{i,c} + \sum_{i=1}^{N} g_{i,c}} \tag{2.5}$$

The mean Dice coefficient across all C classes is:

$$\text{mDice} = \frac{1}{C}\sum_{c=1}^{C} \text{Dice}_c \tag{2.6}$$

Table 2.1: Performance metrics for binary segmentation task. Best results are shown in **bold**, second best are underlined.

Method	**Vanilla UNet**	**UNet11**	**TernausNet**	**LinkNet34**	**UMobilenet**
Mean IoU	0.88	0.90	**0.91**	0.90	0.90
Mean Dice	0.93	**0.95**	**0.95**	0.94	**0.95**
Parameters	9.16M	13.06M	18.56M	24.34M	**8.32M**
FLOPs	55.40G	12.91G	23.21G	7.45G	**6.32G**

The deep-learning methods used were in an end-to-end pipeline, where the segmentation was performed by taking a raw image as input to generate the final segmentation output. By incorporating a network with fewer learnable parameters and modifying the downsampling technique, we were able to improve performance while generating a more efficient network. We observed that UMobileNet performed as well as the state-of-the-art model (at the time of publication) on binary segmentation with a fraction of the parameters required fro the model. However, we ran into challenges with multi-class segmentation, particularly for instrument type segmentation tasks. All U-net type models performed poorly in identifying instruments, with the best model performing at 0.37. A potential reason for this could be the overwhelming presence of background pixels interfering with the segmentation of thin instrument classes. This is particularly true for the more rarely occurring instruments such as Grasping Retractor and Vessel Sealer.

Table 2.2: Performance metrics for instrument parts segmentation with per-class Dice scores (4 classes: background, shaft, wrist, claspers)

Method	**Vanilla UNet**	**UNet11**	**TernausNet**	**LinkNet34**	**UMobilenet**
mIoU	0.71	0.74	**0.75**	0.74	0.73
mDSC	0.78	0.81	**0.82**	0.81	0.80
			Per-class Dice Scores		
Background	0.99	0.99	0.99	0.99	0.99
Shaft	0.82	0.85	**0.86**	0.84	0.83
Wrist	0.69	0.73	**0.74**	0.73	0.73
Claspers	0.63	0.66	**0.68**	0.66	0.65

Table 2.3: Performance metrics for multi-class instruments segmentation task (8 classes). Best results are shown in **bold**, second best are underlined.

Method	**Vanilla UNet**	**UNet11**	**TernausNet**	**LinkNet34**	**UMobilenet**
Mean IoU	0.33	0.36	0.35	**0.37**	**0.37**
Mean DSC	0.35	0.38	0.37	**0.39**	**0.39**
Std IoU	0.12	0.15	0.14	0.15	**0.15**
Std DSC	0.14	0.16	0.15	0.16	**0.17**

While post-processing techniques such as conditional random fields or grab cut [77] could be applied to further improve performance, we wanted to further understand how we could learn from the rich, high-dimensional image data for multi-class segmentation in an end-to-end framework.

2.4 Improving Instrument Type Identification Using Hybrid Transformer-CNN architecture

While convolutional neural networks provide a significant improvement from classical machine learning techniques for image segmentation, there are limitations to their use due to the challenges in deriving global context in images, as they are unable to model long-range dependencies. The computer vision community has recently benefited from the advances in the transformer models [78] proposed by the Natural Language Processing community, and derived a version suitable for images called the Vision Transformer [66]. The key contributing factor is the use of self-attention in transformers, which provides global context to the models [79]. A limitation is, however, the models are compute-intensive, making their use in segmentation tasks for real-time use challenging.

Various approaches have identified techniques to combine the utility of transformers with the advantages of CNNs to develop segmentation models [80–83]. Inspired by these approaches, we develop an efficient hybrid transformer-CNN model for surgical instrument segmentation called SurgSeg-HViT.

The core model architecture for SurgSeg-HViT is presented in Figure 2.5, and is inspired from the U-net architecture. The primary difference here is an encoder that has both convolution and vision transformer components. For the ViT blocks, we make use of the MobileViT model [84], while the CNN components are located within the MFConvGating-SC block. The decoder is a mirror of the encoder, much like in U-Net models, with the cascading decoder structure incorporating multi-scale information. The MFConvGating-SC block uses GSConv [85], which merges standard convolution and depthwise separable convolution by shuffling the channels. The convolution information is combined at different resolutions, as shown in Figure 2.7. The importance placed on both the channel-level features and spatial-level features provides improved feature selection capabilities in the MFConvGating-SC block. A skip connection is added to the block to avoid the vanishing gradient problem.

The feature incentive blocks and the long local block are incorporated based on [81]. The downsampling technique used is Haar wavelet downsampling from [86] instead of

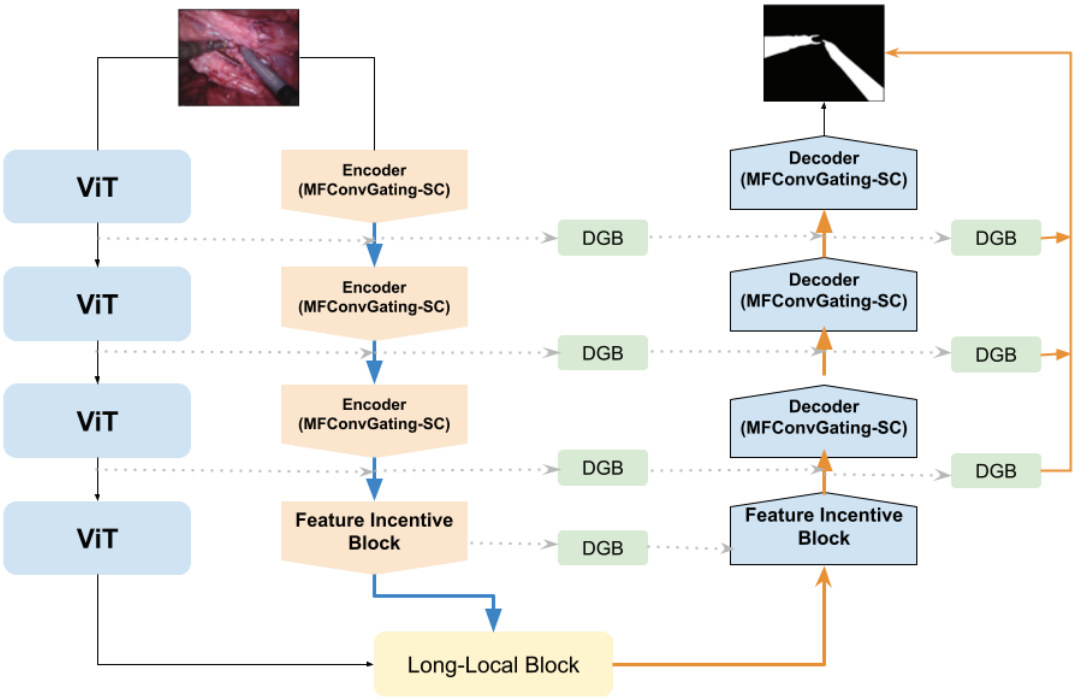


Figure 2.5: SurgSeg-HViT model architecture for a hybrid Transformer-CNN surgical instrument segmentation task

maxpooling, which proves to be more sensitive to capturing high frequency information. This is done by using the Haar Wavelet Transform, where a feature map of the input is split into their frequency components. This is indicated using the blue arrows in Figure 2.5. Similarly, upsampling (indicated in orange) uses the technique from [87], which combines commonly used upsampling techniques to provide refined boundary information. These modules are particularly helpful for surgical segmentation since the edges of structures are poorly defined. Overall, the model architecture for SurgSeg-HViT is based on efficient components from [81,83,85,86], and [87], which makes the model lightweight in the different blocks used. Additionally, the Dynamic Gating Block (DGB) described in Figure 2.6 is used to improve feature fusion by using a dynamic approach to weighting channels.

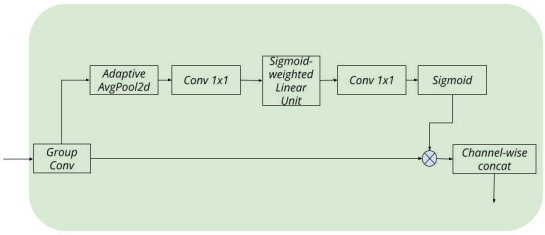


Figure 2.6: Dynamic Gating Block (DGB)

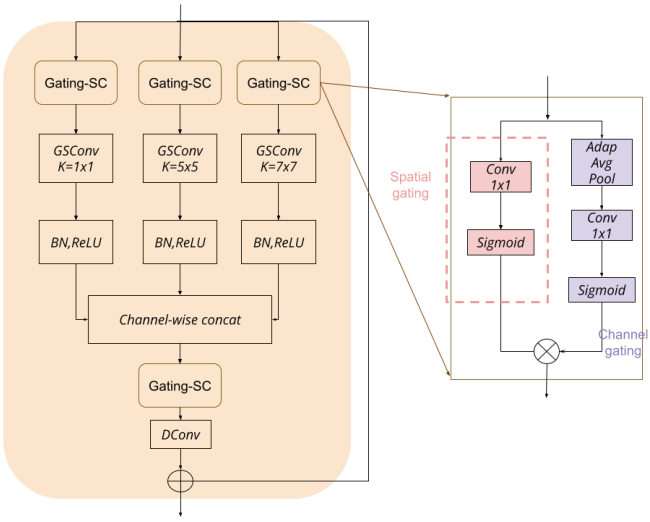


Figure 2.7: MFConvGating-SC Block enables multi-fusion at different resolutions and offers spatial and channel gating for improving model performance.

2.4.1 Analysis

For training, a combination of binary cross-entropy loss, dice loss and Jaccard loss are used, with a heavier weighting of 0.6 to dice loss, while Jaccard and BCE have a weighting of 0.2. We compare SurgSeg-HViT method for surgical instrument type identification with the previously evaluated U-net models using the mean Intesection over Union and mean Dice-Sorenson metrics. The model shows a significant increase in performance for the surgical instrument segmentation task, with an mIoU of 62.74 and mDSC of 63.16.

Table 2.4: Performance on Endovis2017 dataset for instrument type segmentation task

Method	**Vanilla UNet**	**UNet11**	**TernausNet**	**LinkNet34**	**UMobilenet**	**SurgSeg-HViT**
Mean IoU	0.33	0.36	0.35	0.37	0.37	**62.74**
Mean DSC	0.35	0.38	0.37	0.39	0.39	**63.16**
Std IoU	0.12	0.15	0.14	0.15	0.15	0.31
Std DSC	0.14	0.16	0.15	0.16	0.17	0.33

Qualitative results of the different models on the Endovis2017 dataset are provided in Figure 2.8 for Binary segmentation, Figure 2.9 for parts segmentation, and 2.10 for instrument segmentation. We also observed an average inference time of 30.58 frames per second, which enables the use of SurgSeg-HViT in real-time applications. Therefore, through this project, we were able to achieve a much higher performance on surgical instrument type segmentation.

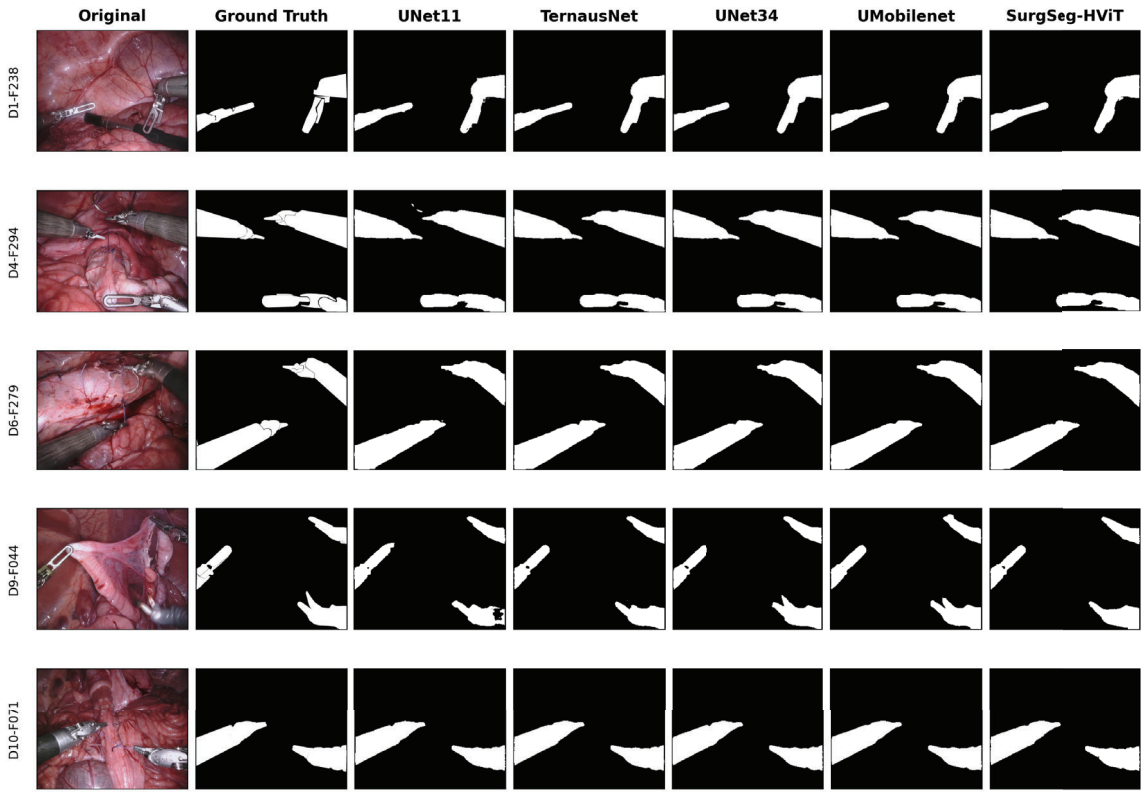


Figure 2.8: Qualitative results on MICCAI EndoVis2017 Dataset for Binary Segmentation

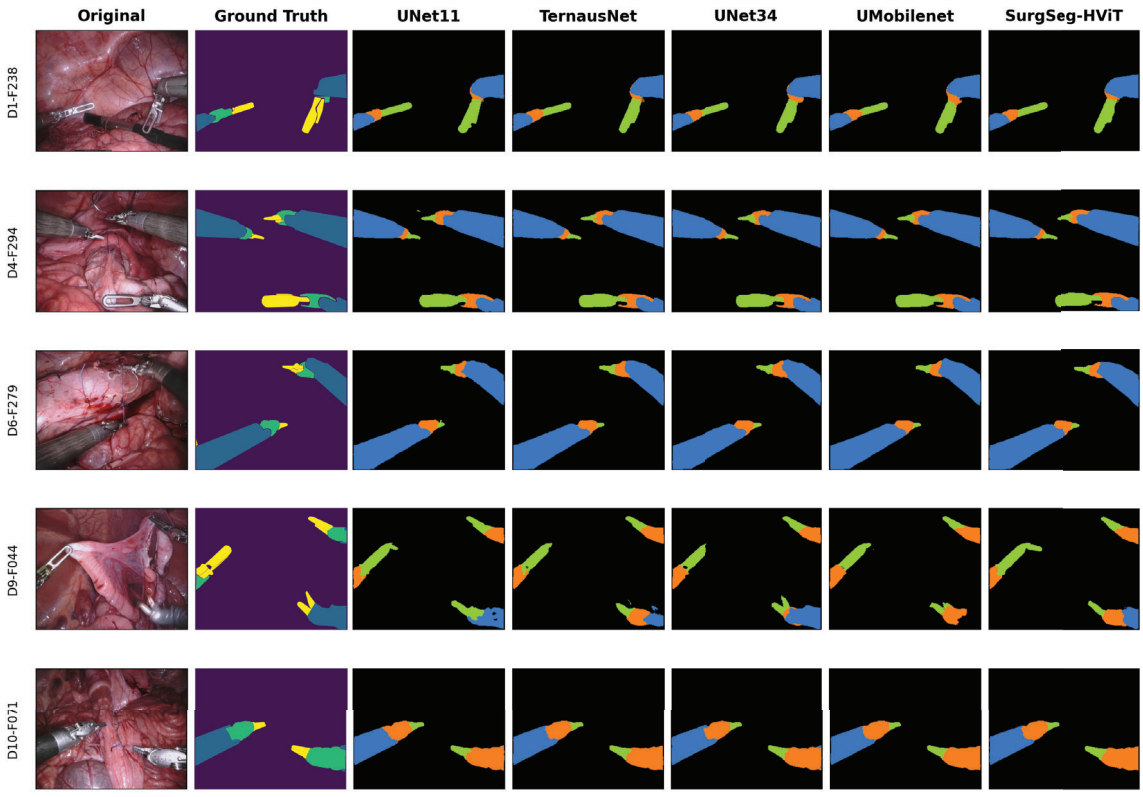


Figure 2.9: Qualitative results on MICCAI EndoVis2017 Dataset for Parts Segmentation

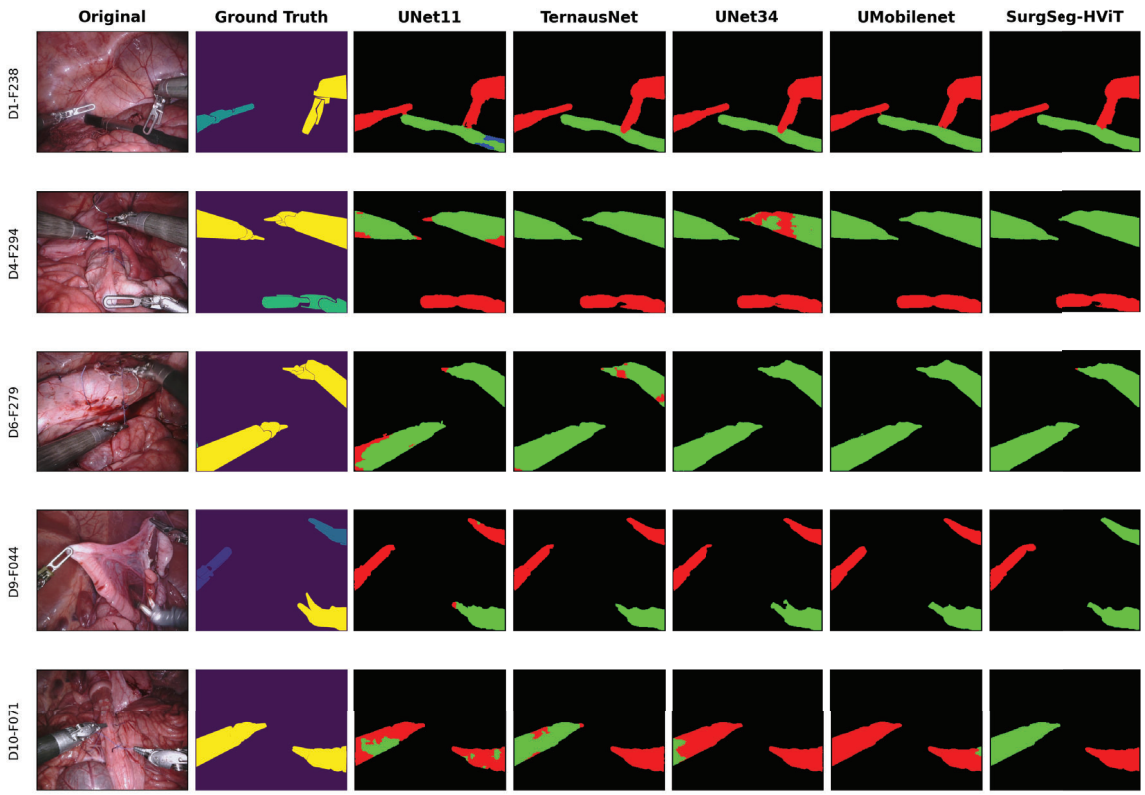


Figure 2.10: Qualitative results on MICCAI EndoVis2017 Dataset for Instrument Segmentation

Chapter 3

DATA EFFICIENT LEARNING FOR SURGICAL VISION

3.1 Background

The modern operating room is a complex and continuously changing environment that requires innovative intraoperative clinical support mechanisms to enable safe procedures. Surgical Data Science [88] has been gaining a lot of traction since the advent of deep neural networks for its ability to process large volumes of data and provide insights about the surgical scene. Considering the safety-critical nature of surgery and surgical robotics, it becomes important to identify approaches for surgical scene analysis that scale efficiently and incorporate domain-specific knowledge simultaneously. As established in the previous chapter, surgical procedures rely on visual output from either endoscopes or microscopes that provide valuable information about patient anatomy and current surgical state, enabling surgeons to safely navigate the surgical workspace.

A large part of the progress in surgical data science has relied on a relatively small volume of data compared to the field of general computer vision. While there is no shortage of data from surgical procedures, generating data annotations has been challenging due to the cost involved with surgeon time and expertise [50]. Additionally, it is essential to manage and process the influx of data without spending enormous amounts of money on compute and storage. There are also concerns around the distribution of data within these datasets due to lack of data involving rare conditions, privacy concerns of patients or the surgical team, high cost of data capture and the variability of conditions (recording equipment, lighting, etc.,) at data collection sites. Therefore, it is important to identify approaches to efficiently process unlabeled data that could potentially reduce the required manual annotation effort, automate some aspects of data usage for model training, and expedite the training cycle.

In chapter 2, we adopted techniques such as data pre-processing, data augmentation [76], and transfer learning [35,69] to optimize the use of available labeled data. Some other issues

with supervised learning techniques is around the inability to handle class imbalance [89] effectively. Surgical datasets have a long-tail distribution of classes where certain classes occur more frequently. This causes a bias towards the majority class and rare but important classes could be ignored. Shifts in data distribution [90] over time could also cause model performance to decrease if corrective measures are not taken. Another challenge is with incremental learning [91] where the model knowledge is continuously updated as new data becomes available. Incorporating efficient model training and data processing frameworks for surgical robotics would contribute to model generalization, reduce computational and storage requirements, and result in systems useful for real world clinical needs.

Towards this goal, we propose two approaches to efficiently train machine learning models for surgical scene analysis through semantic understanding - one with active learning and the other using foundation models. In this chapter, we detail the use of active learning to query data points from a large unlabeled data pool by using uncertainty estimation. In Chapter 4, we will describe an approach based on transfer learning and foundation models for reference-based surgical instrument segmentation.

3.2 Active Learning

3.2.1 Prior Work

Active learning (AL) is an area within machine learning that speeds model training by querying the most informative images for the model from an unlabeled data pool [92]. Active learning uses either (i) Uncertainty-based methods [93–97], (ii) Diversity-based methods [98, 99], (iii) Hybrid methods using a combination of uncertainty and diversity estimation techniques [100, 101], or (iv) Expected model change [102, 103]. These methods could either be applied to a whole image or could select regions of an image to annotate. The uncertainty-based methods estimate the images for which the neural network model is least certain how to label.

While most of the active learning methods have focused on identifying samples for image classification [93–96, 99, 104], the approaches to identify samples for semantic segmentation have been limited [97, 103, 105–109]. Most of these focus on pool-based active learning,

where the most informative annotation samples are chosen from a pool of unlabeled data as opposed to stream-based active learning, where samples are obtained one at a time instead. This stems from deep learning models learning from a mini-batch of labeled data to train more efficiently and avoid local minima [110].

Ensemble methods [95] have been used to estimate uncertainty between model predictions. However, these require a lot of compute and memory. Uncertainty metrics that are commonly used for active selection from image data sets include entropy and least-confident posterior probabilities [111]. Another metric is to choose samples which have the least difference between prediction confidences of the first most probable label and the second most probable label [104]. These can be used directly with deep learning models but have to be implemented carefully to avoid heavy memory demand.

[97] based their work on cost-sensitive active learning and used Monte Carlo Dropout to estimate uncertainty combined with effort estimation from the click patterns during annotation by a human oracle. While the click information from many datasets may not be available, we focus on the region-based study of images from this work. A more interesting development to this paper was from [107] who exploited the multiple views of the same scene to estimate uncertainty of superpixels.

Another development has been the use of reinforcement learning for querying images. Casanova et al. [108] propose an active learning strategy using deep reinforcement learning where a learned policy would select a informative image regions from the unlabeled dataset. They use a Deep Q-Network(DQN) [112] to maximize the *Intersection over Union*(IoU).

In this work, we evaluate the use of uncertainty as the acquisition metric to identify informative surgical image samples to be manually labeled, and the incorporation of a diversity metric to improve the selection of data points.

3.2.2 Methods

Given a segmentation model $f_\theta : X \to Y$ parameterized by θ, we aim to iteratively select the most informative frames from an unlabeled video pool $\mathcal{D}_U$ to expand our labeled dataset $\mathcal{D}_L$. Due to the robust performance of DeepLabV3+ [113] on segmentation benchmarks, we

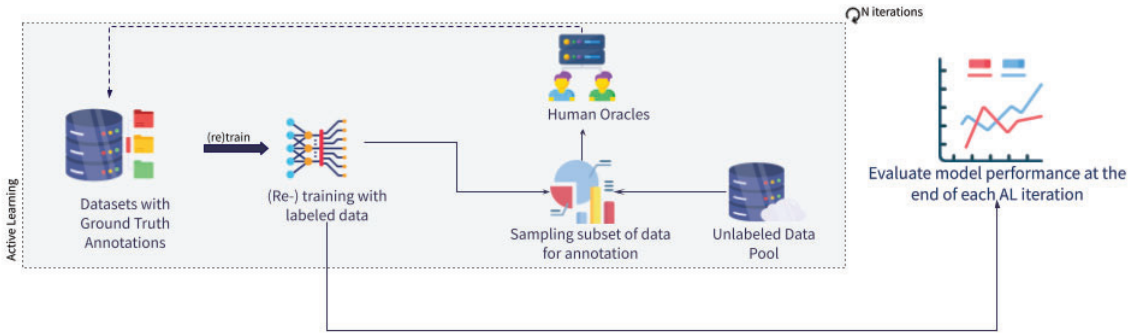


Figure 3.1: Overview of Active Learning for Surgical Instrument Segmentation. A pre-trained machine learning model is used to query frames from an unlabeled data pool by using a selection criterion.

use it as our f_θ of choice for identifying surgical tools from endoscopic images. The backbone is initialized with weights from a model that was pretrained on data from the ImageNet classification task. The rest of the layers are initialized with Kaiming initialization [114].

We use two base uncertainty metrics to sample images. The first is using Monte Carlo Dropout (MCD), proposed by [115]. Dropout [116] is a type of regularization added to neural networks to prevent overfitting. Including spatial dropout [117] in convolutional neural networks implies we drop entire channels with a selected probability value of $p_{dropout}$. Due to the nature of their design where nodes can create bottlenecks for different sections, we use dropout at inference time to emulate the behavior of ensembles without adding to the computational cost of training and inference. Predictions are generated using forward inference for N iterations and uncertainty of the model is estimated by calculating entropy over the N predictions for each sample. The Monte Carlo Dropout method is given by:

$$S_i = -\sum_{n}\sum_{i=1}^{M} p(y_i^n|x_i;\theta_{SEG}) \log p_{MC}(y_i^n|x_i;\theta_{SEG}) \tag{3.1}$$

Here,

$$p_{MC}(y_i^n|x_i;\theta_{SEG}) = \frac{1}{D}\sum_{d=1}^{D} p_{SM}(y_i^n|x_i;\theta_{SEG}^d) \tag{3.2}$$

Here, D denotes the number of forward iterations to run dropout at test time, M is the total number of examples, x is the input, y is the output, θ_{SEG}^d is the segmentation model parameters. Depending the uncertainty value, represented by S_i, the top-k frames will be selected in vanilla MCD. The other base uncertainty metric we use is the entropy of class predictions.

$$S_i = -\sum_{n}\sum_{i=1}^{M} p(y_i^n|x_i;\theta_{SEG}) \log p(y_i^n|x_i;\theta_{SEG}) \tag{3.3}$$

We study the use of vanilla MCD and Entropy, along with variations including region-based estimates and combinations with diversity methods such as coreset. Coreset [99] is an approach used to represent datapoints by assigning a centroid and computing the distance of data points to the centroid. By using a hyperparameter to define the budget for images with similar representations, we can limit the number of samples selected from each cluster. Weighted uncertainty metrics combine randomization with top-N image selection to provide a top-k subset. Here, the images the model is most uncertain about are weighted more heavily than images where the model confidence for predictions are high. The variants of region-based inference consider the uncertainty of 32x32 patches in each image instead of estimating the uncertainty of the whole image. Hybrid entropy fuses uncertainty metrics (entropy and MCD) with coreset.

3.2.3 Analysis

We evaluate our approach on the publicly available 2017 Robotics Instrument Segmentation dataset [33] The dataset publicly shared 1800 images with labels. We use 1500 images for training and validate on 300 images. Our baseline method is random selection, where images are selected at random for each iteration and added to the training pool. Results in Fig. 3.3 show the performance of all models for each active learning iteration. The modified uncertainty-based models outform the vanilla uncertainty acquisition functions. The diversity-based Coreset methods performs the least favorably. This could be due to the

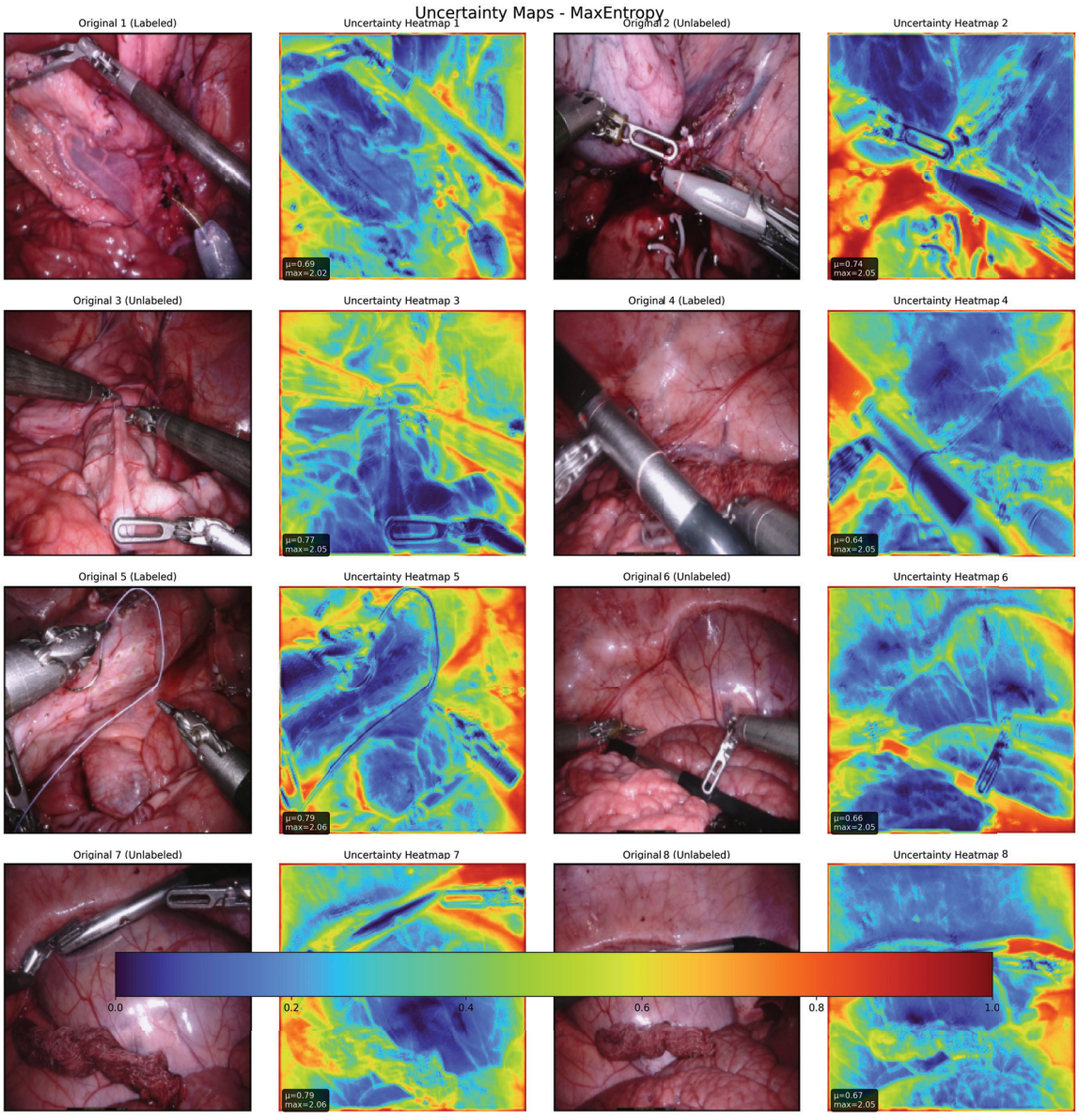


Figure 3.2: Qualitative results using Active Learning showing the uncertainty in the top batch of images selected for labeling by the model.

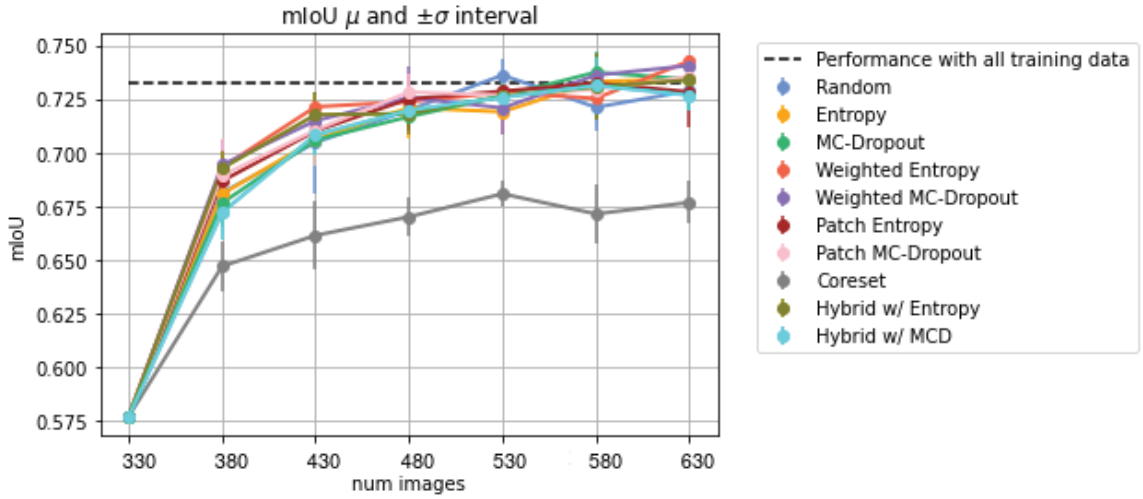


Figure 3.3: Active Learning Results: Overall performance of the model is achieved by using a maximum of 630 images (instead of 1500) for training a model with active learning.

selection of data points that are distinctly different from each other. Since the dataset is relatively small, the model is potentially unable to extract enough information for representing the entire dataset. Table 3.1 shows the amount of time taken to select images in each iteration of the active learning cycle.

Table 3.1: Time and mIoU values for different active selection strategies. The estimated time taken to calculate uncertainty is indicated in seconds for 1 image.

Active Learning approach	**mIoU**	**Time(in secs)**
Baseline (random)	0.728	0.00019
Entropy	0.7344	0.04076
Monte Carlo Dropout	0.73408	0.11579

3.3 Balanced Expected Gradient Length with Density-Weighted Selection

Following our work, additional studies on active learning for surgical instrument segmentation were proposed [118, 119]. A limitation similar to our analysis in diversity sampling impacted the performance of sampled data. The coreset from our hybrid work and kNN+ diversity sampling technique used in [118] spread the images as far as possible. A more refined approach would be to use a weighting approach to sampling diverse datapoints to ensure sufficient samples of uncertain images are included in the annotation pool.

To address this limitation, we propose a Balanced Expected Gradient Length with Density-Weighted Selection (BEGL-DWS) framework for surgical instrument segmentation. BEGL-DWS has three key components: (i) class-balanced uncertainty quantification accounts for the varying sizes and frequencies of different surgical tools, (ii) pixel-level gradient length estimation focusing on instrument boundaries and articulation points, and (iii) density-weighted batch selection to ensure temporal and visual diversity across surgical phases. The complete procedure is outlined in Algorithm 1.

Surgical scenes exhibit significant class imbalance, with the background occupying the majority of pixels while individual instruments—particularly thin tools like sutures or needles—represent only small regions. Moreover, the frequency of different instruments varies substantially across surgical phases, with some specialized tools appearing only briefly during specific procedural steps. To address this challenge, we compute adaptive class-specific weights based on the current labeled set:

$$\alpha_c = (\text{freq}_c + \varepsilon)^{-\beta} \tag{3.4}$$

where freq_c represents the relative frequency of instrument class c in $\mathcal{D}_L$, ε is a small constant for numerical stability, and $\beta \in [0.5, 1]$ controls the strength of balancing. The weights are normalized such that $\sum c = 1^C \alpha_c = 1$.

For each unlabeled surgical frame $X \in \mathcal{D}_U$, we compute the expected gradient length as a measure of model uncertainty, which is particularly effective for capturing uncertainty around instrument boundaries, articulation joints, and regions with motion blur or occlusion. The gradient length provides a geometry-aware uncertainty estimate that naturally

Algorithm 1: BEGL-DWS: Balanced Expected Gradient Length with Density-Weighted Selection

Input : Segmentation model f_θ with features z and predictions p; unlabeled set $\mathcal{D}_U$; labeled set $\mathcal{D}_L$; batch size B; AL rounds N

Input : Top-pixel ratio $\rho \in (0, 1)$; class weight power $\beta \in [0.5, 1]$; k-neighbors k; density weight $\lambda \in [0, 1]$

Output: Trained model f_θ after N rounds

Initialize: Train f_θ on initial $\mathcal{D}_L$
for $n \leftarrow 1$ **to** N **do**
  freq $\leftarrow ComputeClassFreq(\mathcal{D}_L)$
  for $c \in \{1, \ldots, C\}$ **do**
    $\alpha_c \leftarrow (\text{freq}[c] + \varepsilon)^{-\beta}$
  end
  $\alpha \leftarrow \alpha / \sum_c \alpha_c$
  $\mathcal{Q} \leftarrow \emptyset$; $\mathcal{E} \leftarrow \emptyset$
  foreach $X \in \mathcal{D}_U$ **do**
    $q_X, \phi_X \leftarrow ComputeBEGL(X, f_\theta, \alpha, \rho)$
    $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{q_X\}$
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{\phi_X\}$
  end
  $\mathcal{S} \leftarrow DensityWeightedSelect(\mathcal{E}, \mathcal{Q}, B, k, \lambda)$
  $\mathcal{D}_L \leftarrow \mathcal{D}_L \cup SurgeonLabel(\mathcal{S})$
  $\mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \mathcal{S}$
  $f_\theta \leftarrow Train(f_\theta, \mathcal{D}_L)$
end
return f_θ

emphasizes the complex contours and thin structures characteristic of surgical instruments. We first identify the most uncertain pixels by selecting the top ρ proportion of pixels based on prediction entropy:

$$H_{i,j} = -\sum_{c=1}^{C} p_{i,j,c} \log p_{i,j,c} \tag{3.5}$$

These high-entropy pixels typically concentrate along instrument boundaries, at occlusion interfaces, and in regions where instruments overlap or interact with tissue. For these selected pixels $\mathcal{P}\rho$, we compute the class-balanced expected gradient length:

$$q_X = \frac{1}{|\mathcal{P}\rho|} \sum_{(i,j)\in\mathcal{P}_\rho} \sum c = 1^C \alpha_c \cdot p_{i,j,c} \cdot |\nabla_\theta \ell(f_\theta(X)_{i,j}, c)|_2 \tag{3.6}$$

where ℓ denotes the segmentation loss and the gradient is computed with respect to the model parameters θ.

Finally, to ensure diversity in the selected batch and avoid annotating redundant frames from similar surgical contexts, we employ a density-weighted selection strategy. For each candidate frame with embedding ϕ_i and BEGL score q_i, we compute its density in the feature space using k-nearest neighbors:

$$d_i = \frac{1}{k} \sum_{j\in\mathcal{N}_k(i)} \exp\left(-\frac{|\phi_i - \phi_j|_2^2}{2\sigma^2}\right) \tag{3.7}$$

where $\mathcal{N}_k(i)$ represents the k-nearest neighbors of frame i and σ is a bandwidth parameter. High density values indicate frames from common surgical phases or frequently occurring instrument configurations, while low density values suggest unique surgical scenarios or rare instrument combinations.

The acquisition score constitutes a hybrid approach, combining uncertainty and diversity:

$$s_i = q_i \cdot (1 - \lambda \cdot d_i) \tag{3.8}$$

where $\lambda \in [0, 1]$ controls the trade-off between uncertainty and diversity. This scoring mechanism ensures that selected frames span different surgical phases, instrument combinations, and challenging scenarios, maximizing the information gained from each annotation round.

We apply this learning technique for both Endovis2017 and NeuroID datasets. We seed the models with 440 images, test on 1200 images, validate on 360 images, and treat the rest as unlabeled pool. This approach is compared with a subset of methods and the results are shown in Figure 3.4. We can see a distinct difference in performance compared to the cluster of metrics we observed in Figure 3.3.

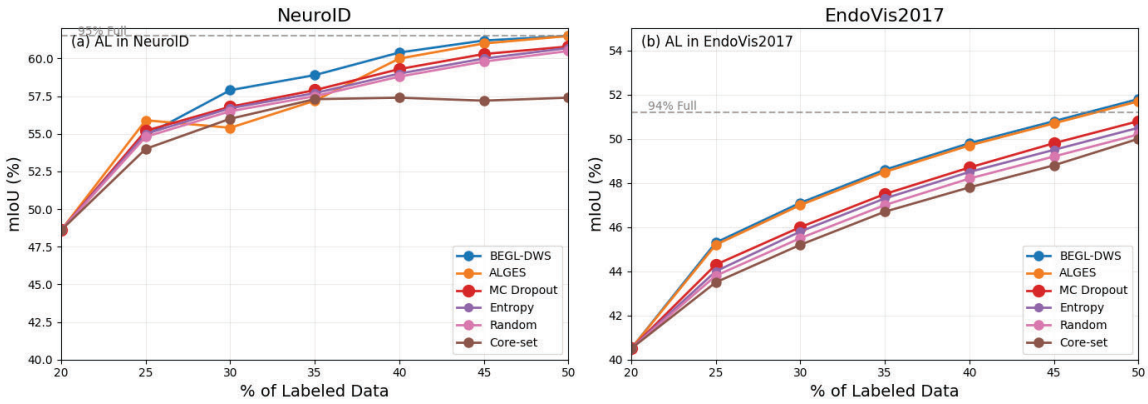


Figure 3.4: Results of BEGL-DWS on EndoVis2017 and NeuroID datasets

3.4 Evaluation of Noisy Labels on Active Learning

The ability to quickly and correctly annotate large datasets is crucial to supervised machine learning. Machine Learning (ML) models may be used in high-stakes decisions, but the time and cost of expert labor make obtaining labeled data expensive. The process of labeling data for an ML task navigates a trade-off between the ability to collect data quickly and the effect of the resulting dataset quality on model performance.

An annotation system that augments the capabilities of human annotators with Artificial Intelligence (AI) could improve this labeling process if the Human-AI team annotates with higher efficiency and accuracy than either the human or AI alone. In terms of efficiency, annotation tools in which the human evaluates and modifies annotations generated by an AI can decrease human workload and make the annotation process faster [120–122].

However, in terms of accuracy of the resulting model, the effect is unclear. Although in an ideal Human-AI team, the human and AI would catch each other's mistakes to achieve higher performance together, work in Explainable AI (XAI) suggests that increasing AI-assistance often does not lead to better performance than relying on just the AI or human alone [123–125]. The tendency of people in Human-AI teams to rely on the AI's prediction even when it is wrong, a phenomenon known as overreliance, could affect the final dataset when using AI assistants for labeling data. Although models can be robust to some noise [126], changing

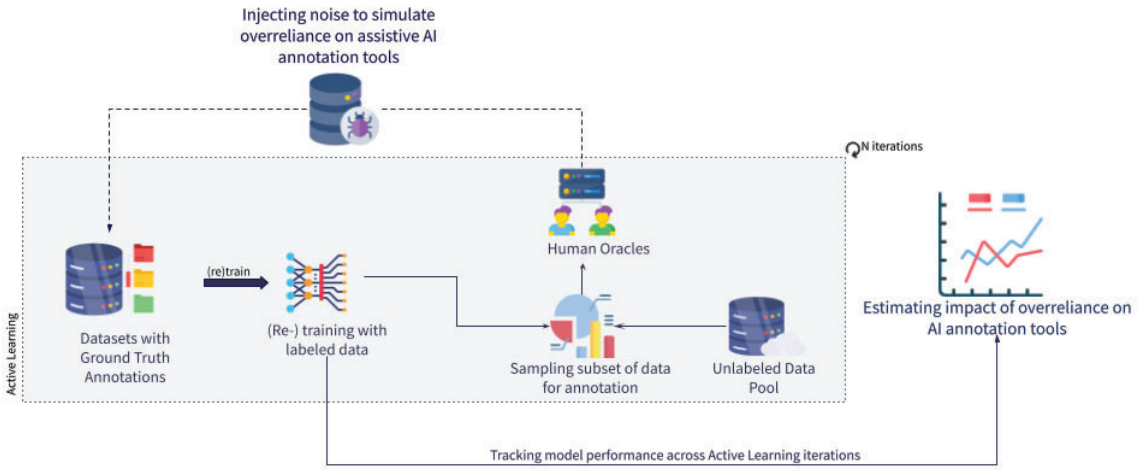


Figure 3.5: We explore the downstream effects of a human oracle relying on AI-assisted annotation tools by simulating noisy labels

dataset quality could also affect the final model's performance.

In our work, we aim to understand how overreliance within Human-AI data annotation teams affects the resulting model. As the underlying goal is to reduce the burden of data labeling, we choose the ML approach of active learning (AL) because it prioritizes labeling of difficult-to-learn data points and attempts to reduce the amount of data needed for training the model. Active learning (AL), specifically pool-based active learning, is a learning protocol that samples the most informative data points for a model by querying from an unlabeled data pool [34]. The objective is to maximize a combination of task performance and data efficiency at the annotation level. Studies evaluating active learning for classification and segmentation validate the advantages of using a well-defined querying policy over randomly sampling data for annotation [94, 101, 108, 127].

Although overreliance is a known and pervasive problem, it is difficult to study its general effects for two reasons. First, the negative impact of overreliance also depends on how often the AI is wrong. Overreliance is not an issue with an oracle AI, but could cause harmful outputs with a low-quality AI predictor. Therefore, to make our results applicable to a wide variety of situations, we synthetically alter the quality of the AI's output by injecting

noise into the labels. Second, the amount of overreliance varies greatly depending on the task type and presentation [123–125, 128]. We simulate various levels of overreliance by changing the percentage of labels taken from human labels and those taken from the AI. Using the MedMNIST [129] and CIFAR100 [130] datasets, we will simulate the effects of these various overreliance conditions on the final model's performance using two noise injection techniques.

Unlike prior work, this approach combines the techniques developed in the active learning space with the human behavioral challenges explored in XAI studies. Work at this intersection is valuable because data-annotation schemes have a different set of motivating factors than other human-in-the-loop systems. Unlike the decision making scenarios of many XAI studies, data annotation methods must consider not only accuracy, but also speed, efficiency, and expense.

3.4.1 Methods

We conducted a series of experiments to examine the impact of human over-reliance on AI-generated noisy labels in an Active Learning scenario for both a medical task and a generic coarse-grained computer vision task. Specifically, we created a synthetic experimental framework in which we generated "AI predictions" using the Noisy Completely at Random (NCAR) algorithm with controllable noise ratios. We then simulated human reliance on these synthetic AI-generated labels in an active learning setting by randomly sampling from a dataset that contained ground truth labels. By adjusting the proportion of labels obtained from human annotators (i.e. the ground truth labels) versus those generated by AI (i.e. generated by NCAR), we simulated varying levels of over-reliance and assess their effect.

Active Learning: In a pool-based Active Learning (AL) framework, there is often a large volume of unlabeled data $\mathcal{D}_U$, and an oracle, usually a human, that annotates assigned data points x_u. Iteratively, AL aims to sample the most informative samples, x_u, based on a pre-determined budget from $\mathcal{D}_U$. The sampled data points are sent to the oracle, labeled as x_i, and subsequently added to the labeled set $\mathcal{D}_L$. Active learning uses an acquisition

function, which is based on a desired uncertainty and/or diversity metric, to choose samples from an unlabeled pool. In a general form, all the informative samples selected by the algorithm are added to the labeled set for model training in the next iteration is given by $D_{i+1} = D_i + x_i$. We use uncertainty metrics to sample data points under the assumption that the most informative examples are the ones that the model is least certain about.

We use three acquisition functions to compare active learning methodologies - Random Sampling, Entropy Sampling and Bayesian Active Learning by Disagreement (BALD). In the Random Sampling Strategy, data points are naively selected at random from the unlabeled dataset $\mathcal{D}_U$. Since the likelihood of all samples being equally informative is low, we use this approach as our baseline. For entropy sampling, we estimate uncertainty by calculating uncertainty from predicted probabilities as given in equation 3.3.

For the BALD approach, we use Monte Carlo Dropout (MCD) to estimate uncertainty. Dropout [116] is a type of regularization added to neural networks to prevent overfitting. Due to the nature of their design where nodes can create bottlenecks for different sections, we use dropout at inference time to emulate the behavior of ensembles without adding to the computational cost of training and inference. BALD determines data points that maximize the mutual information between the model parameters and model output. We estimate mutual information in combination with MCD, given by equation 3.1.

Noisy Label Generation We use two statistical noisy label generation methods: Noisy Completely at Random (NCAR) and Noisy at Random (NAR) [131] to simulate AI predictions. A dataset of n samples can be denoted by $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where x_i is the i-th input feature vector and y_i is the corresponding true label. The set of possible classes is denoted by $C = c_1, c_2, \ldots, c_K$, where K is the total number of classes. To simulate label noise using the NCAR method, a fixed proportion of the true labels are randomly flipped to incorrect labels. Specifically, a parameter $p \in [0, 1]$ represents the probability of flipping each label. Then, each sample i, is randomly flipped its true label y_i to a new label y_i' uniformly at random from the $K-1$ possible incorrect labels, excluding the true label y_i, so NCAR produces symmetrical noise. NAR on the other hand produces asymmetrical label noise, where the occurrence of the noise is tied to the sample's true class such that a noise transition matrix that holds the information of how each class is confused with every other

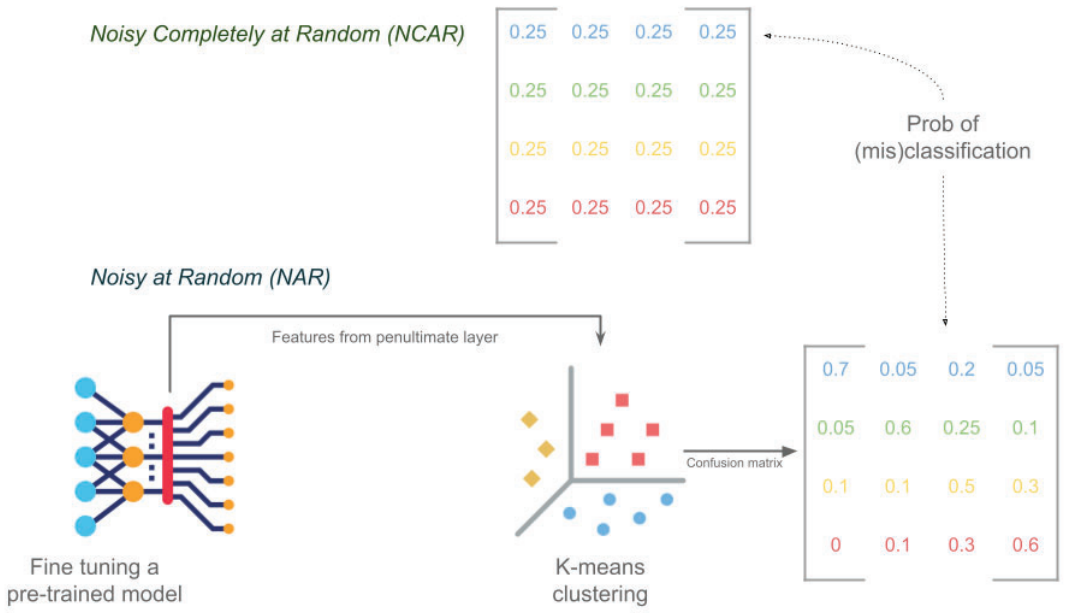


Figure 3.6: Noisy injection uses information from two techniques for noisy label identification – Noisy Completely at Random (NCAR) and Noisy at Random (NAR). NCAR simulates mislabeling where all classes are equally likely to be reassigned. NAR uses a weighted transition matrix for classes.

class in the format of a confusion matrix. The noise transition matrix can be computed by training a k-means clustering on the training data to measure how each true label matches with the clusters. More formally, for NAR, each sample i, is flipped its true label y_i to a new label y_i' by using the probability values given in the noise transition matrix so y_i' is sampled asymmetrically. Figure 3.6 shows the noise injection process.

3.4.2 Experiments and Results

Using OrganMNIST from MedMNIST [129] as our preliminary dataset of choice, we simulate the effects of the various overreliance conditions on the final model's performance. We train multiple active learning models, each using a different combination of the following attributes. We vary the reliance rate ϕ of a human annotator and the noise in our labels

among {0, 0.2, 0.5, 0.8}, and we vary the active learning strategy among random, entropy, and BALD. To inject noise, we use the NCAR and NAR methods for OrganMNIST and the NCAR method for CIFAR-100 (following the similar trend that NCAR and NAR do not yield substantial differences for OrganMNIST, and to save some computational time in our experiments.).

We utilize a modified ResNet-18 convolutional neural network which accommodates the small sizes of the images in our dataset by using fewer max pooling operations. For each active learning strategy, we sample 3% of the samples from our unlabeled dataset for 5 iterations. We use cross-entropy loss for all scenarios, and re-initialize the network after every active learning iteration to train from scratch. We then compare all experiments to the baseline, which is trained once on a data subset with the same size covered by the active-learning runs. At each step, we train the model on the data subset for 100 epochs or until a training accuracy of 0.92 is reached.

OrganMNIST is in the medical domain, which has images of 28x28 dimensions each. We choose a medical dataset specifically as labeling medical data often requires expert human annotators so could benefit from AI-assisted data labeling. CIFAR-100 has slightly higher-resolution color photos with 32x32x3 inputs and is a larger dataset, albeit with fewer samples per class. Initially, we planned to conduct experiments using only medical datasets and chose to use ISIC. However, we pivoted from experimenting on ISIC due to confounding errors caused by the corruption of the label base by actual noise labels, duplicates in the test set, and significant class imbalance [132]. This led us to realize that our assumption of "oracle human labels" falls short on ISIC specifically, where the results seem to get better by random noise (potentially fixing the already corrupted labels along the way)

3.4.3 Discussion

Although overreliance is a known and pervasive problem, it is difficult to study its general effects for two reasons (1) the negative impact of overreliance also depends on how often the AI is wrong, (2) the amount of overreliance varies greatly depending on the task type and presentation [123–125,128]. To make our results applicable to a wide variety of situations, we

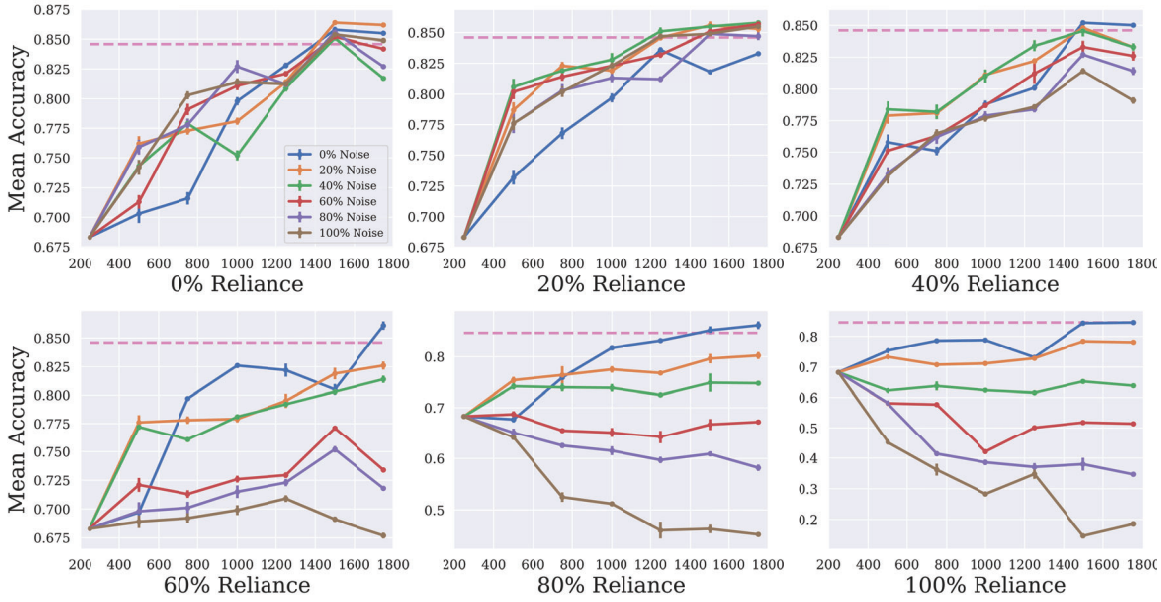


Figure 3.7: We evaluate the downstream effects of a human oracle relying on AI-assisted annotation tools by simulating noisy labels using NAR. The performance of a model trained on the entire OrganMNIST dataset is included in all the experiment results for comparison.

synthetically alter the quality of the AI's output by injecting noise into the labels. Through our results on OrganAMNIST, we observe that for a higher degree of simulated over-reliance there is a correlation with a decrease in performance. The effect is more prominent for higher levels of noise, implying that the model performance within AI-assisted annotation tools is of importance. In cases where only a small amount of noisy labels were passed (cases with 20% reliance or 20% noise), the model performance can have a small boost. We believe this could be a result of the regularizing effect from noisy labels indicated in [133]. In these scenarios, incorporating AI predictions could help improve efficiency in data annotations through a small increase in performance, as long as over-reliance is minimal.

Furthermore, while we expected the sparsity of the Monte Carlo Dropout approaches to result in better model performance than other sampling techniques, we did not have the chance to dive deep into the impact of model performance as a function of the chosen acquisition function. Conducting further studies on the effect of the uncertainty acquisi-

tion functions and contrasting performance with other diversity-based or hybrid acquisition functions may open new research opportunities. Finally, translating the study design to other tasks such as segmentation, object detection, and pose estimation would be interesting avenues of exploration.

3.4.4 Acknowledgements

The work on noisy labels in active learning was jointly done in collaboration with Madeleine Grunde-McLaughin, Wisdom O. Ikezogwo, and Mehmet Saygin Seyfioglu.

Chapter 4

TOWARDS GENERALIZING PERFORMANCE FOR SURGICAL SCENE UNDERSTANDING USING SELF-SUPERVISED LEARNING

Active learning helps reach improved performance with fewer data samples but it still needs a large volume of annotated data initially to reliably generalize for unseen surgical conditions and new types of surgeries. Active learning also has the potential to skew a model based on the type of data points chosen for annotation, missing on valuable information from data points that are not accounted for due to sampling budget constraints.

Self-supervised learning (SSL) is an active area of research in machine learning and computer vision that has the potential to substantially reduce the dependency on manually annotated data, which is expensive and time-consuming to generate in surgical domains. SSL overcomes the need for large volumes of annotated data by harnessing the inherent structure and relationships within the data to provide supervisory signals for extracting representations. By pre-training on the generated representations, SSL can potentially enable training with significantly fewer labeled data points. A simplified difference between supervised and self-supervised learning is presented in Fig. 4.1.

Furthermore, depending on the type of surgical task and data, surgical data annotation suffers from inter-annotator variability. This makes the data annotation process tedious since labels have to obtained from multiple annotators and techniques to resolve disagreements have not been standardized. Having to consistently generate annotations with a human-in-the-loop can be limiting with respect to the demands of continual learning, which is resolved by techniques in self-supervised learning. SSL also allows scaling models by pre-training on large volumes of unlabeled data. The pre-trained model can be used for finetuning a variety of downstream tasks using fewer labeled data points.

The advances in generative AI and foundation models have provided a new class of pretrained models that have been trained on internet-scale data. Transfer learning enables

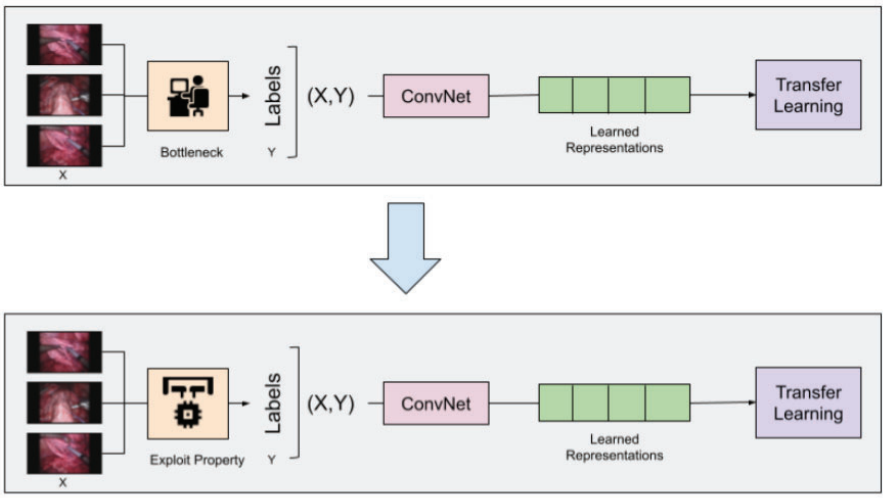


Figure 4.1: An overview of the self-supervised learning structure. Instead of relying only on annotated data, the use of pre-trained models and transfer learning enables improved generalization abilities.

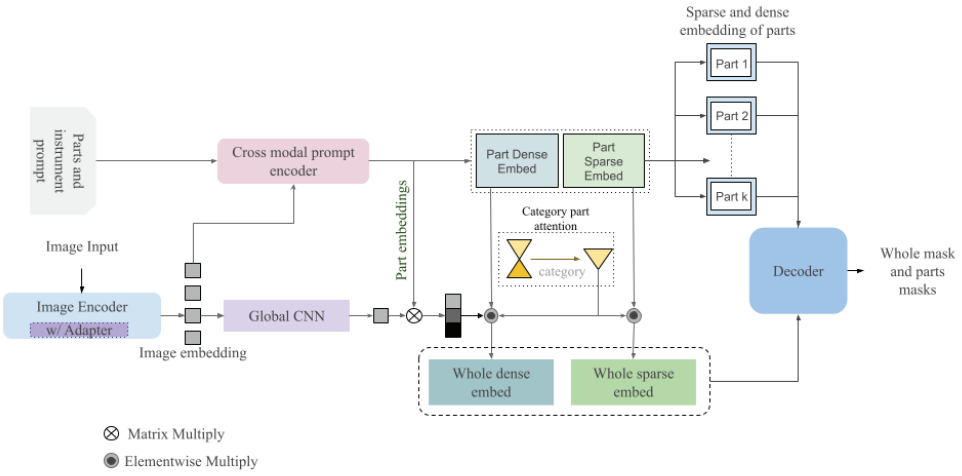


Figure 4.2: Model Architecture for performing reference-based surgical instrument segmentation.

improved domain adaptation by learning domain-invariant features from the large pool of unlabeled data and provides robustness to noise in annotations. This is due to the reliance of foundation models on representing high-level features (often associated with human vision) and contextual information, which captures versatile data representations in the learned embeddings. This is unlike supervised learning scenarios where correctness of the annotation labels is prioritized and the need for experts to be involved is paramount. Additionally, the possibility of the model results being skewed due to the type of images selected for annotation is lesser in transfer learning.

4.1 Background

Self-supervised learning has experienced rapid traction within the computer vision community in recent years [134]. The potential to solve problems without bearing the cost of annotating a large volume of data is enticing. SSL is an approach where a proxy loss is minimized, which is used as a supervisory signal for capturing information that is inherently represented in the data. This can be established by (1) defining a auxiliary task such as image rotation [135], arranging image patches [136], colorization [137], reconstructing

image patches [138] and geometric transformations [139], or (2) contrastive methods [140] that focus on how data is represented in the embedding space by generating positive and negative pairs of samples that are processed by a discriminative loss function to generate a training signal, or (3) Cluster-based and distillation-based methods [141, 142] focused on clustering latent image representations in the embedding space.

During surgical procedures, surgical video is typically recorded for review or for educational purposes. The storage of raw video data is memory-intensive but the information presented in each case could be valuable for downstream learning. A majority of studies in the surgical vision community have relied on labeled data, typically much smaller in size than the datasets used by the broader computer vision community [143]. This has limitations in accounting for variations in patient anatomy, skills of the operating surgeon, devices used for image acquisition, types of surgical instruments, different surgical phases and clinical workflows, and patient demographics.

Yengera et al. [144] use the prediction of remaining time in a surgical video as a proxy task to represent information relevant to surgical phase recognition. The hypothesis is that the discriminative features learned in the embedding space will implicitly capture information about the steps completed thus far and the steps left until a procedure is completed. Similarly, [145] demonstrate the use of temporal coherence and compare three methods for pre-training using SSL. Imprecise kinematic information was used by [146] to do 3D pose estimation of surgical instruments, utilizing a reconstruction loss. Motion2Vec [147] learns a motion-centric representation of surgical demonstration from videos by grouping them into categories (actions, sub-goals, options) using metric learning in a Siamese network.

Depth estimation, in monocular and stereo endoscopy, has been of growing interest in the surgical vision community. A number of publications have focused efforts on depth estimation using self-supervised learning [148–153]. Relatively fewer research teams have directly exploited information by prioritizing image frames in surgical vision for other applications. A research team from Germany used self-supervised learning to establish rotation-equivariant keypoint detection and invariant description of endoscopic scenes using an approach termed RIDE [154]. The authors use image pairs to which known homographies are applied. The learned representations were tested on datasets for surgical tissue tracking

and endoscopic stereo correspondence matching [155].

4.1.1 Reference-based surgical instrument segmentation

Accurately segmenting surgical instruments from endoscopic video is a foundational capability for intraoperative visual intelligence, enabling reliable AR overlays, context-aware assistance, and safety monitoring in robot-assisted minimally invasive surgery [156]. Over the past decade, community benchmarks, particularly the EndoVis challenges, have standardized problem formulations and evaluation, catalyzing progress on binary, type-level, and part-level segmentation of instruments [33,157]. Early deep models adapted general-purpose segmentation architectures, such as U-Net variants described earlier in this thesis, to explored residual backbones and multi-scale design for improved accuracy and runtime [74]. Subsequent work pushed toward clinical practicality with real-time instance segmentation and attention mechanisms under challenging endoscopic conditions [158].

Recently, segmentation foundation models have shifted the landscape. The Segment Anything Model (SAM) demonstrated strong zero-shot capabilities with promptable masks [159], followed by SAM 2, which extends to images and videos with a streaming memory design [160]. Domain transfer to medicine has required significant adaptation as shown by MedSAMm which achieves broad cross-modality performance via large-scale medical finetuning [161]. However, endoscopy remains a elusive for open-world models to segment accurately due to availability of fewer samples in the original training data, and due to factors such as specularities, smoke, blood, motion blur, telephoto optics, and intricate forceps-scissor-jaw geometries. In surgical vision specifically, class-promptable adaptations such as SurgicalSAM efficiently tune SAM to instrument categories [162], while part-aware prompting (SP-SAM) leverages collaborative part-to-whole cues to better capture fine end-effector boundaries [163]. These studies suggest that combining promptable foundations with surgical priors (category taxonomy, kinematic parts, tool functional zones) is key for robust MIS segmentation.

Parallel advances in referring image segmentation offer an alternative path, where the use of text description specifies the target region instead of fixed labels. Transfer learning us-

ing CLIP-based transfer aligns text-to-pixel for open-vocabulary masks [164]. The study by Yang et al. [165] demonstrates language-aware visual encoding that fuses tokens throughout a transformer encoder, and transformer-only designs improve long-range reasoning between language and vision [166]. Beyond single-object expressions, generalized referring experession segmentation supports multi-/no-target expressions and richer relations [167]. Most relevant to surgical tools, multi-granularity reference expressive sefmentation (MRES) explicitly extends from object-level to part-level referring masks [168], similar to the part-based compositionality of articulated instruments (shaft, wrist, jaws). Bringing refeering image segmentation into minimally invasive surgery could enable natural-language-driven segmentation of specific instruments and substructures. For example we could provide prompts such as "the distal cutting edge of the curved scissors touching the cystic duct", addressing cases where class-only prompts are insufficiently precise.

However, two technical challenges remain. The first is estimating the boundary more accurately. Given the critical need for safe guidance of surgical instruments, it is critical for the thin structures, such as wires, jaws, tips, to not be over-smoothed by region-based objectives. Therefore, it is salient to include boundary-centric metrics and losses to complement the region IoU [169, 170]. Additionally, open-vocabulary grounding must account for domain shift and ambiguity in terms used for referring instruments. Therefore, for the purpose of MIS, our proposed methodology decouples be the promptable segmenter adapted to surgery from the referring module that resolves granular, part-aware linguistic queries, and maps them to discriminative visual prototypes.

We study referring surgical instrument segmentation propose an approach that uses a part-aware, promptable foundation. Our approach develops a Part-aware Referring SAM pipeline that (i) adapts SAM to surgical videos, (ii) injects part-level priors via collaborative prompts and hierarchical decoding to sharpen end-effector boundaries, and (iii) uses a lightweight CLIP-initialized head for open-vocabulary and text-driven selection of instruments and sub-parts. We further incorporate boundary-sensitive training and evaluation to quantify improvements at the thin, safety-critical interfaces. In comprehensive experiments on Endovis2017, we show that our proposed part-aware referring improves both mask quality over the state-of-the-art part-aware segmentation model.

4.2 Method

Given an endoscopic image frame, we predict a dense instrument-part map, Y, with labels $\{0 : \text{background}, 1 : \text{shaft}, 2 : \text{wrist}, 3 : \text{jaws}\}$. We build a part-head over frozen Segment Anything (SAM) image embeddings and expose two text-to-class gating mechanisms, a training-free CLIP gate and a learned text projection, to bias predictions toward the referred part. We condition inference on a referring phrase that includes the type of instrument and the part of the instrument, such as "needle driver tip". Additionally, we utilize the adapter in [171] to refine SAM for a surgical instrument segmentation task.

For each image we compute a dense SAM embedding with 256 channels using a frozen SAM image encoder. A lightweight head turns the embedding into per-pixel class logits using the following operations:

$$Z = ReLU(GN(\text{Conv}_{1\times1}(E))) \in \mathbb{R} \times C \times h \times w \tag{4.1}$$

$$L = \text{Conv}_{1\times1}(Z) \in \mathbb{R}^{4\times h\times w}, \tag{4.2}$$

$$L_{up} = \text{bilinear_upsampling}(L) \tag{4.3}$$

The two 1×1 projections with GroupNorm and ReLU preserves the localization of SAM while keeping compute and parameters minimal.

4.2.1 Referring Interface for part segmentation

We use two techniques to map a free-form phrase, t, to a target part, c - (i) Mapping lexicons (using synonyms for parts), and (ii) Encode a frozen CLIP-text encoder and compare it to canonical prompts using cosine similarity.

4.2.2 Text-to-Class Gating

Beyond selecting the class (c), we treat text as a class-prior bias added to the per-pixel logits. Given $w \in \mathbb{R}^4$ from CLIP similarities, we add a spatially uniform bias

$$\tilde{L}_{up} = L_{up} + \alpha \cdot 1_{H,W}. \tag{4.4}$$

with α=2.0 by default at inference and $1_{H,W}$ an all-ones map broadcast over (H, W).

We freeze CLIP and learn a single linear layer W of 4 dimensions for the classes. W maps the normalized text embedding to 4-way class biases and the relationship is given by

$$\tilde{L}_{up} = L_{up} + \alpha \cdot b(t) 1_{H,W}. \tag{4.5}$$

where $b(t) = \phi(t)W$. The linear adapter better aligns domain lexicon with the surgical part taxonomy at negligible compute cost.

4.2.3 Prompt Diversity Schedule

To expose the learned gate to varied phrasing, we employ a simple scheduler over synonyms: (i) random-per-batch (default): draw a random synonym for a random part each batch; (ii) resample-per-epoch: at epoch start, build a balanced pool with at least K unique synonyms per part and iterate cyclically. This prevents overfitting to a single canonical wording and stabilizes the text loss.

4.2.4 Losses

Let Y be the ground truth and $\tilde{L}_{up}$ be the gated logits.

Segmentation: We use cross-entropy and dice loss, similar to those described in chapter 2 and 3. Here,

$$\mathcal{L}_{\text{seg}} \;=\; \text{CE}\Big(\tilde{L}_{up}, Y\Big) \;+\; \mathcal{L}_{\text{dice}}\Big(\tilde{L}_{up}, Y\Big)\,, \tag{4.6}$$

where Dice is computed class-wise and averaged over parts.

Prompt supervision: We sample a prompt-label pair $(t, c^\star)$ from the synonym pool and impose

$$\mathcal{L}_{\text{text}} \;=\; \text{CE}(b(t),\, c^\star)\,. \tag{4.7}$$

Therfore,

$$\mathcal{L}_{total} \;=\; \mathcal{L}_{\text{seg}} \;+\; \lambda\, \mathcal{L}_{\text{text}} \tag{4.8}$$

4.3 Experiments

For training and inference, we use a frozen Segment Anything Model. The images are resized to 640, with random horizontal flip (p=0.5). We train the part head and the text projection

with AdamW (lr 1×10^{-4}, wd 1×10^{-4}), mixed precision, gradient clipping (1.0), batch size 8, for 25 epochs. Validation uses plain logits (no gating) to report intrinsic segmentation quality.

4.4 Analysis

We use SAM for image features and CLIP for text embeddings. The learned adapter adds a single $D\times4$ linear component, where D is the resolution. We run our experiment using two NVIDIA GPUs (Titan X and Titan RTX). We implement the architecture in PyTorch.

We summarize the results from performing reference-based segmentation on the EndoVis2017 dataset in the table below.

Table 4.1: Reference-based Surgical Instrument Segmentation on the EndoVis2017 Dataset.

Method	Mean IoU	Instrument Categories						
		BF	PF	LND	VS	GR	MCS	UP
TernausNet	35.27	13.45	12.39	20.51	5.97	1.08	1.00	16.76
ISINet	55.62	38.70	38.50	50.09	27.43	2.10	28.72	12.56
TrackAnything (1 Point)	54.90	47.59	28.71	43.27	**82.75**	63.10	66.46	55.54
TP-SIS	63.37	66.42	45.46	75.20	73.44	29.95	44.02	34.67
Mask2Former + SAM	66.21	66.84	**55.36**	83.29	73.52	26.24	36.26	45.34
SurgicalSAM	69.94	68.30	51.77	75.52	68.24	57.63	86.95	**60.80**
SP-SAM	73.94	68.89	53.16	83.80	73.20	72.40	84.91	41.05
Ours	**75.82**	**70.45**	55.28	**85.67**	74.85	**74.18**	**87.23**	43.10

Chapter 5

HUMAN-CENTRIC SURGICAL ROBOT FOCUSED ON SUB-TASK AUTOMATION

In robot-assisted surgery, the most widely used systems, such as the da Vinci Surgical Robot (Intuitive Surgical Inc. [20]), are predominantly teleoperated. Surgeons control these robots from a console, performing maneuvers with a back-drivable manipulator or joystick that are replicated by a patient-side robot. The robot executes commands in near real-time and does not independently perform surgical tasks or sub-tasks. Surgeons are responsible for data analysis and planning during both pre-operative and intraoperative stages, as well as for all decisions and motions that guide the robot. The introduction of autonomous robotic assistance may improve surgical safety, speed, and efficiency. Automating repetitive tasks can enhance workflow by leveraging the dexterity and precision of surgical robots [172].

Some commercially available robotic systems, such as the ROBODOC (currently sold as the TSolution-One by Think Surgical) [173] used for orthopaedic surgery, and the NeuroMate [174] robot used for electrode implantation for deep brain stimulation (DBS) and stereotactic neurosurgical procedures, employ autonomy under the supervisory guidance of surgeons. ROBODOC integrates preoperative planning using 3D Computed Tomography (CT) scans with real-time feedback for bone preparation, implant alignment, and positioning for hip and knee arthroplasty. NeuroMate provides pre-operative planning features that enables precise targeting of trajectories and a steady robotic arm with virtual fixtures, which is teleoperated by the surgeon to mitigate accidental movements. These robotics demonstrate the benefits of automation in advancing surgical techniques and illustrate how autonomous operations can augment the surgical workflow.

Yang et al. [36] repurposed the standard J3016 set by the Society of Automotive Engineers to define levels of autonomy for self-driving cars [175] and established a framework for six levels of autonomy for surgical robotics. Level 0 implies no autonomy in the clinical set-

ting, where a robot is entirely teleoperated by the surgeon. Here, the surgeon performs tasks ranging from pre-operative planning, generating surgical options to perform the procedure, selecting the appropriate option, and finally executing the surgical procedure. This has been an established approach to integrating robots into the operating room. On the other hand, Level 5 denotes full automation, where a surgical platform operates autonomously to complete the aforementioned tasks, a level that does not include a surgeon in the loop. There is no current research that employs principles of full automation. It is preferable to develop surgeon-in-the-loop systems to ensure safe operation of surgical robots. In this regard, Levels 1 to 4 have surgeon oversight while robots execute tasks. Level 1 includes robot assistance where the surgeon maintains continuous control of robot operations, and assistance is primarily to provide cognitive or mechanical assistance. The ROBODOC is a level 1 system, as the surgeon drives the robot milling operation until a virtual fixture is reached, similar to NeuroMate. Task autonomy enables robots to execute specific tasks autonomously, even without a comprehensive understanding of the surgical workflow. This is covered under level 2, where the robot is not capable of replanning and relies on the surgeon to initiate task execution. Level 3 is conditional autonomy; the robot has an improved understanding of surgical workflow and creates a plan, which it then replans and executes following the surgeon's approval. The most desired level of autonomy sought by researchers is currently Level 4. Here, the robot has the ability to utilize preoperative data to devise a surgical plan, execute the plan, and account for changes using intraoperative analysis and replan if necessary. Given the dynamic nature of surgical procedures, the capacity to make modifications is salient. A limitation of this structure is that, unlike self-driving cars, surgical procedures involve more than one operator in the surgical workspace, providing oversight during the overall procedure.

While [36] accounts for the type of decision-making (or lack thereof) performed by a robot, the role of a surgeon can vary within each level depending on the type of surgery, context of the surgical task and the type of pathology. [176] and [177] together provide a user-centric classification of autonomy by codifying the type of interactions between a surgeon and the surgical robot. We extend these and propose six level of autonomy based on how control is effected - (1) *Direct control*: where the surgeon drives decision-making and

executes the procedure using a surgical robot, typically through teleoperation or manual guidance, (2) *Supervised autonomy*: where the surgeon makes high-level decisions on tasks that the robot would autonomously execute under their supervision, (3) *Shared-control*: where the surgeon and the robot make decisions cooperatively, (4) *Traded control*: where the procedure is decomposed into sub-tasks that are interchangeably completed by the surgeon and the robot but the robot still functions under the oversight of the surgeon, (5) *Restricted unconditional autonomy*: where robots are capable of identifying and dealing more rapidly than human surgeons with designated emergencies, or (6) *Full autonomy*: where the robot makes autonomous decisions without the supervision of a surgeon. Direct control is similar to what occurs in Level 0. Shared control is typically viewed as level 1, as the surgical systems designed thus far have only been designed to guide surgeons' movements. Supervised autonomy and traded control can encompass levels 2, 3, and 4. The restricted unconditional autonomy state is situated at the intersection of levels 4 and 5. Full autonomy corresponds to level 5.

Successful Human-Robot Interaction within an operating room relies on interactions not just between the robot and the surgeon but also with the surgical team. The surgical team plays a crucial role in the success of surgeries, to the extent that effective communication between teams can significantly impact the outcome of a surgery, either positively or negatively [178] [179]. With the exception of the SARAS project [180] and SACHETS [30], research in surgical automation has predominantly centered on tasks performed by the surgeon and has not accounted for the presence of another user in the surgical workspace. To address this gap, this chapter describe a benchmarking system to develop collaborative autonomy. The framework utilizes the surgical scene understanding components described in the previous chapters towards autonomously performs assistive suction on a tabletop experiment setup, mimicking the functions of a surgical first assistant and providing support alongside the surgeon during robot-assisted interventions.

5.1 Autonomy in Surgical Robotics

Building on the benefits observed in robotic minimally invasive surgery through teleoperation, there is growing interest in introducing autonomous execution of surgical sub-tasks

Table 5.1: Summary of autonomous surgical task and sub-task studies and the perceived level of autonomy upon successful *in vivo* implementation

Autonomous Surgical Task/Sub-task	References	Level of autonomy [23] [36]
Autonomous Debridement	[182] [186] [187] [188] [189] [190]	Level 3
Tumor Ablation	[191] [192] [193] [194] [195]	Level 2
Tissue Manipulation	[196] [197] [198] [199] [200] [201] [202] [195]	Level 2
Tensioning and Cutting	[203] [204] [205] [29] [206]	Level 2
Suturing	[207] [208] [209] [181] [210] [211] [212] [213] [214] [215]	Level 2 & 3
Autonomous Camera Control	[183] [184] [216] [217] [218] [219]	Level 2
Fundamental Laparoscopic Surgery Training Tasks	[185] [220] [221] [206]	Level 2
Autonomous Suction	[30] [186] [222] [223]	Level 2

to support the surgical workflow [179]. Automating repetitive and tedious tasks could potentially reduce the workload for users, ensure consistency in treatment, and decrease the duration of the surgery.

There has been considerable effort in automating sub-tasks performed by surgeons through shared control, traded control, or supervised autonomy. Some common applications of autonomy have included automating surgical subtasks, such as suturing [181], autonomous debridement [182], autonomous control of the endoscope [183,184], and a Fundamentals of Laparoscopic Surgery (FLS) task - peg transfer [185].

Krupe et al. [224] proposed autonomous 3-D positioning of surgical instruments using visual servoing. The authors utilized light-emitting diodes on the instrument tip and incorporated an instrument holder to project laser dots, thereby localizing the instrument with respect to the patient's anatomy. This is an example of a level 1 system. Other level 1 systems could encompass enhanced visualization [225] and haptic feedback [226].

Autonomous endoscope control has been an active area of research aimed at alleviating the cognitive demands placed on surgeons to simultaneously control surgical instruments and the endoscope. [183,184,219] have focused on using an instrument-centric approach and controlling the position of the endoscope based on the position of other surgical instruments used during the procedure. Authors of [218] study the visual attention of surgeons and demonstrate that predicting surgical actions can guide the choice of visual attention for

autonomous endoscope control.

In [185], Hwang et al. develop a *superhuman* peg transfer where the autonomous framework performs better than an experienced surgical resident. The system incorporates an enhanced calibration technique to enhance the measurement accuracy of the end effector and a nonlinear trajectory optimization module to facilitate bilateral handover and complete the task. Other projects from the same group focused efforts on autonomous debridement of viscoelastic tissue phantoms [182, 188, 189].

A particularly interesting project from Johns Hopkins University is the Smart Tissue Autonomous Robot, commonly referred to as STAR [207, 207, 227]. TThe project performs an autonomous intestinal anastomosis procedure under a supervised autonomy model for collaboration. The surgeon selects from automatically generated surgical plans, and the STAR robot autonomously executes a range of tasks, including tissue tracking, compensating for breathing motion, selecting a plan, and replanning if required, as well as predicting tool collisions, among others. The overall system architecture evidences the potential benefit that can be realized by integrating multiple sub-components of autonomous surgical control.

We include some of the studies on autonomous debridement, suturing, autonomous endoscope control, knot tying, cutting, tissue manipulation, and tumor ablation that we reviewed regarding surgical autonomy in Table 5.1. Achieving level 4 autonomy requires reliable models of the patient and a definition of the task to be performed. While various components of developing reliable patient models, registering patient position and anatomy intraoperatively, processes to define steps in a surgical task, methods to execute surgical tasks, and safety strategies have been independently researched, there have been no research prototypes exploring high autonomy yet.

5.2 Metrics to Estimate Uncertainty and Communication of Competence Within Human-Robot Teams for Surgery

With robot-assisted minimally invasive surgical systems installed in over 8,000 locations (including, but not limited to, the da Vinci Surgical Robot) [228], understanding their impact in the operating room and on OR team proficiency becomes crucial. While the community

has explored potential research directions in human-robot interaction for surgery [14] [229], analysis around how we evaluate interactions involving a surgical robot has been lacking.

Additionally, developments in the field of machine learning have led to the creation of new tools for surgical robotics research, as summarized in [230], [172], and [23]. In conventional surgeries, surgeons can self-assess their approach to a procedure at every stage by considering the operating conditions and accounting for their own abilities, limitations, and the specific goal of the given procedure. As more autonomous capabilities are integrated into the operating room, it is vital to study how surgeon-robot teams interact under varying levels of autonomy [36].

Humans interacting with robots tend to form conceptual models of robot behavior. By surveying the perspectives of surgeons participating in surgical robotics research, Aaltonen and Wahlstrom [231] present methods to address user experience challenges and highlight the need for an extensive study of trust, safety, ergonomics, and security. Developing assistive robotic systems for surgery requires an understanding of how trust between a surgeon and the robot develops, especially in situations of uncertainty and vulnerability. In [232], the analysis from real surgeries highlights the importance of interfaces that help reduce uncertainties in a surgeon's decision-making. Towards this goal, our work proposes a metrology-based framework to center experiences of the surgeon and develop effective mechanisms for communicating robot competence by deriving from surveying surgeons, evaluating results from past user studies in surgeon-robot teams at the University of Washington [191,233–236], and by reviewing existing the human-robot interaction metrics [237–240] used in the surgical robotics literature. This effort is motivated by the need for a common basis for evaluating surgeon-robot interactions, which will subsequently enable advances in collaborative surgical robots [28, 182, 241].

We provide descriptions of the uncertainty and attributes categories identified from participatory discussions grounded in the surgical robotics literature, which are provided below:

Alignment of uncertainty and performance: This is used to estimate the correlation between the estimated variance of a model output and the true observations from performing a specific sub-task [242].

Risk-averse reward: A robot's performance must be evaluated based on both correct and incorrect actions, as well as the robot's inaction. The robot's confidence estimates can be used to determine whether performing a task is more or less beneficial than not performing it and deferring to the surgeon or surgical team instead [243].

Uncertainty reduction: It is important to estimate scenarios where the robot is uncertain about its predictions or actions. Methods such as online active learning can be used to identify uncertain scenarios where the robot actions have to be fine-tuned [244].

Reliability: This measures the success of a robot in achieving a specific goal, whether for an overall task or for intermediate subtasks. One could also use time to estimate reliability, as well as the mean time between failures and the mean time to failure [245] [182].

Communicability: The level of abstraction or amount of detail being communicated, the length of information being provided, its simplicity, and comprehensive nature will all determine whether the established communication channels work as intended or not [246].

Medium of Communication: The interfaces that we could use to communicate with surgeons is aplenty. It is essential to determine whether this would be textual, image-based (such as saliency maps or 3D reconstructions), Augmented/Virtual Reality headsets, Haptic Feedback [241], or through other mechanisms [231].

Information communicated: This would include information regarding new data points observed by the robot, potential estimates of failures in subsequent actions, or the probability of success versus failure, etc.

5.3 A framework for benchmarking sub-task collaboration in surgical robotics

5.3.1 Prior Work

Systems where surgeons and robots collaborate to complete tasks have been studied through a traded control framework. In these settings, the surgeon and robot take turns completing a surgical task that is positioned as a shared goal. [247] uses Hidden Markov Models for context modeling and temporal curve averaging based on Dynamic Time Warping to learn motion from demonstrated data for pin transfer and suturing tasks. With context modeling, the transition phases between surgeon-operated and autonomous operations are established.

The primary surgeon operates robotic systems from a remote surgical console. Teleoperated robots filter tremors, scale surgeon movements for precision, and enhance dexterity and ergonomics. The first assistant, positioned near the patient, provides essential intraoperative support. In both robotic and conventional procedures, the first assistant maintains a clear surgical field, manages irrigation, assists with tissue manipulation and retraction, controls bleeding, and handles surgical instruments. First assistants play a critical role by anticipating the surgeon's requirements, coordinating with the surgical team, and facilitating efficient and safe procedures [248]. Therefore, it is important to ensure the first assistant has a manageable workload [249, 250].

Sirasala et al. [251] propose a collaborative approach in which a scrub nurse hands surgical instruments to surgeons. The CISOBOT from the University of Malaga [252] and HuBot from Imperial College London [253] presented conceptual frameworks for assistance to surgeons and demonstrated the performance of their algorithms for shared control to solve Hand-Assisted Laparoscopic Surgeries (HALS) problems and for the completion of a peg transfer task and simulated robotics myomectomy, respectively. Both studies showed improved performance in the efficiency of the procedures under autonomous assistance. [254] studied the teleoperation, supervised control, traded control, and full autonomy collaboration models for segmenting a simulated tumor site. The authors found that using a robot offers improved task accuracy and consistency in segmenting boundaries, and shared control between a human and a robot decreases overall task completion time.

Among the tasks performed by the first assistant, surgical suction and irrigation is a repetitive task, but one that plays a critical role in providing the surgeon a clear field of view. Unlike other surgical scenarios that can be grouped into dedicated surgical phases, the accumulation of blood in the surgical site can occur rapidly and unexpectedly. Excessive bleeding (a.k.a hemorrhaging) and the resulting obstruction of the surgical field have been associated with increased surgical risks [255, 256]. Suction, in particular, plays a pivotal role in controlling hemorrhaging, thereby enhancing the efficacy of the surgery. Efficient, skilled assistants who synchronize with the surgeon and the surgical team, however, can take hundreds of cases to develop [257]. This becomes even more challenging in resource-constrained settings where first assistants may not be readily available. A study by Martinez et al. [258]

studied transferring control of suction-irrigation to surgeons to compensate for the interruptions. While the authors found improved times for surgeon-operated suction, cognitive workload demands and surgeon fatigue were important factors that were not evaluated.

Despite the critical nature of suction, the automation of suction has received little attention in the surgical robotics literature compared to other surgical subtasks, such as suturing and debridement. [223] showed a proof-of-concept setup for autonomous suction, where the authors detected the contour of a region with blood and tested a trajectory planning algorithm to reach the site for suction. The Advanced Robotics and Controls research group (ARCLab) at the University of California, San Diego, has developed an autonomous suction system aimed at achieving hemostasis, a process that stops bleeding from a wound [222] [259]. The authors of [259] developed a probabilistic model to locate the potential source of the bleed and used tracking across frames to generate the desired trajectory for providing suction. While these approaches use instrument segmentation to detect and remove pixels belonging to an instrument from each frame, they do not account for any dynamically moving instruments in a surgical scene.

Barragon et al. [30] propose an autonomous suction tool to reduce the cognitive demands placed on surgeons to control suction tools during laparoscopic surgery. Users would perform a suturing task while the suction tool was either operated through teleoperation by the users themselves or functioned autonomously. The authors use a Fully Convolutional Network with a VGG-16 backbone to perform semantic segmentation of the scene, identifying locations where a pool of blood would accumulate. The suction tool followed a straight-line trajectory from the base position to a target position where the simulated bleeding was identified. The study found that the autonomous modality resulted in users completing their tasks faster than when suction had to be teleoperatively controlled. A limitation of this work is the need for surgeons/users to monitor and adapt to the motion of an autonomous instrument in the surgical workspace.

Project SARAS (Smart Autonomous Robotic Assistant Surgeon) [180] is another study that explores the automation of tasks typically performed by bedside assistants. The SARAS solo-surgery platform cooperatively performs a surgical peg transfer task, where the surgeon performs the task until the ring is handed to the robot. The robot completed the task by

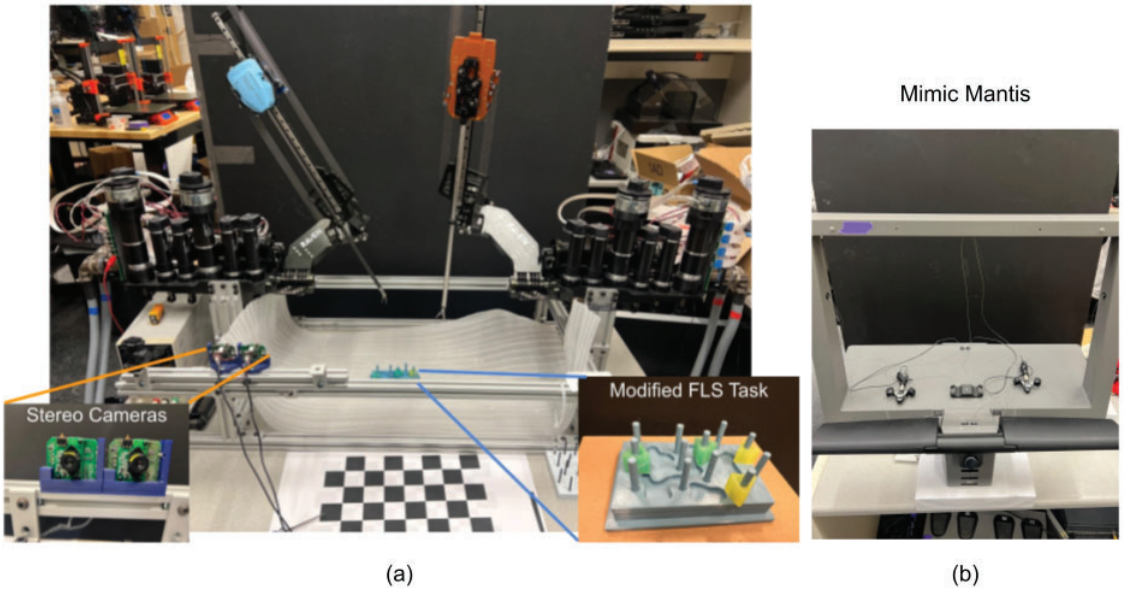


Figure 5.1: Experimental setup of the human-robot collaborative experiment setup. A user will *operate* a laparoscopic grasper instrument to perform peg transfer. This will be controlled by hand, and access to the taskbed will be via a port on a cardboard to simulate remote center of motion. (a) The RAVEN II surgical robot will work as the surgical assistant alongside the user to perform suction on a modified FLS task base. (b) Mimic Mantis will be used to teleoperate the robot. Teleoperation will be performed by a member of the study team.

inserting the ring into a peg. The project proposes using action recognition, a supervisory controller, and a controller to integrate trajectory optimization [260, 261]. The proof-of-concept systems developed by the SARAS project evidence the benefits of automating subtasks performed by a surgical assistant. While performing cooperative tasks where the human and robot work together to complete the same task over a period of time, tasks like suction cannot be performed under a shared control framework. Additionally, studies such as [262] demonstrate that autonomously completing surgical tasks in parallel results in a reduction of at least 40% in the completion time of surgical subtasks.

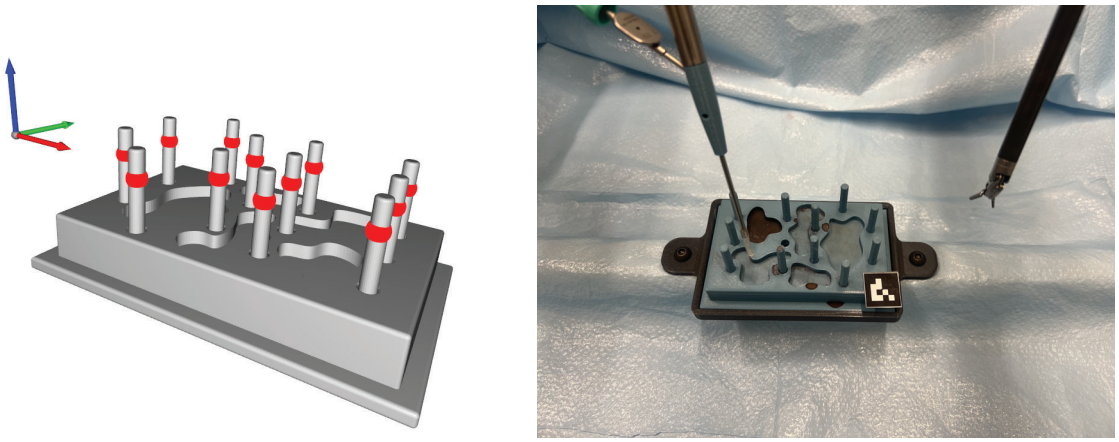
To address this gap, we propose using an autonomous surgical robot that provides suction assistance, capable of navigating a surgical scene with dynamic obstacles, such as other surgical instruments. The proposed system is of Level 2 automation, integrating a type of shared control with respect to the surgeon and supervised autonomy with respect to the first assistant. With the goal of standardizing user-centric descriptions of autonomy, we refer to the interactions established through a project of this nature as *cooperative autonomy*.

5.3.2 Cooperative Autonomy Study Framework

Surgeons plan and execute surgical tasks by augmenting knowledge about patient anatomy and surgical technique with common sense reasoning and implicit knowledge about objects and associated physical properties. For a surgical robot to perform autonomously under dynamic conditions, it requires a combination of surgical scene understanding and some a priori knowledge about the environment it is located in.

Fundamentals of Laparoscopic Surgery is a training curriculum followed by surgeons in training to learn the physiology and technical skills required for basic laparoscopic surgery. It was developed by the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) in 2004 [263]. The curriculum provides a list of distinct and specific tasks that are relevant to laparoscopic surgery, including peg transfer, pattern cutting, endoloop placement/ligating loop, extracorporeal suturing, and intracorporeal suturing. The Fundamentals of Laparoscopic Surgery peg transfer task [264] was chosen due to its importance in surgical training for robotic surgery, its practicality for dry-lab experiments, and its widespread use as a benchmark task in surgical robotics research.

Our framework uses a modified task board for peg transfer to simulate a collaborative goal setup, introducing cavities to emulate blood pooling. A prototype of the 3D printed can be seen in Fig. 5.2. The system setup includes a stereo camera directed at the surgical scene as shown in Fig. 5.1(a). For evaluating collaborative autonomy, the study requires a human subject, typically a surgeon, to move pegs based on a pre-defined pattern provided for each experimental run. In parallel, the surgical robot performs suction in one of two modes: (i) in Mode 1, the robot is operated through manual teleoperation by a member of the study



(a) Task Board Model

(b) Printed Task Board

Figure 5.2: We design a modified peg-transfer task with cavities at the base of the board. This board is a tabletop experiment testbed to benchmark autonomy for surgeon and assistant tasks performed in parallel.

team, and (ii) Mode 2 engages the robot in providing autonomous suction. To teleoperate the surgical robot, we use a Mimic Mantis device (as shown in Fig. 5.1(b)) for positioning the suction tool. We assume that the use of any 3 Degrees-of-freedom (DoF) leader device should yield similar results, considering the device will be operated by a member of the study team. The autonomous mode operates by detecting the presence of blood pooling and surgical instruments using surgical scene analysis as described in the previous chapters. A separate model is trained for the segmentation of blood in cavities, which serves as the goal region for performing suction. To provide reliable suction, we use a mixed-pattern path generation as defined in [265] to identify control points on the contour of the segmented region and follow an optimal trajectory between points to achieve complete coverage.

The overall system is developed by bringing together different components built on perception, motion planning, and control modules, including instrument tracking for surgical scene understanding, control point generation [266], and anytime planning [267] to generate collision-free trajectories for the robotic arm. We impose constraints to limit the velocity

of the suction tool and maintain a minimum safety distance between the surgeon and robot tools to avoid collisions. An ideal user study will involve 8 trials of peg transfer by the user, with 4 trials dedicated to suction being performed in mode 1 and 4 trials of suction via mode 2. The sequences to perform each trial will last a maximum of 4 minutes.

The framework is beneficial in evaluating two hypotheses from the data collected using the cameras and robot kinematic data, which are as follows:

- **H1. (Surgical Efficiency)** Automated Suction provides more efficient removal of blood compared to manual teleoperation.

- **H2. (Surgeon Workload)** Automation of suction reduces the workload of the primary surgeon.

Counterbalancing measures such as randomizing the order in which participants work with the baseline (teleoperation) versus the proposed algorithm for human-centered dynamic obstacle avoidance will mitigate order effects (familiarity, novelty, learning, fatigue, and habituation) [268].

5.3.3 Stereo vision and calibration

While visually perceiving a scene, analyzing the semantic information, as well as geometric knowledge of the scene, ensures safe execution of robot actions. We perform a camera calibration procedure to map pixel coordinates of the task board to x and y position of the robot end-effector using a checkerboard of known dimensions using OpenCV. To determine the location of the surgical instrument in the camera frame, we estimate the relationship between the robot base frame and the camera frame. This relationship is established using a checkerboard pattern placed at a known location from the base of the surgical robot. Using the findChessBoardCorners function in OpenCV, we detect the location of corner points and map it to the known 3D location of the points with respect to the robot base frame. By registering the robot manipulator to the camera frame, we transform the location of objects from pixel space into 3D coordinates for the robot to access. All transformations between coordinate frames are represented with homogeneous transformation matrices.

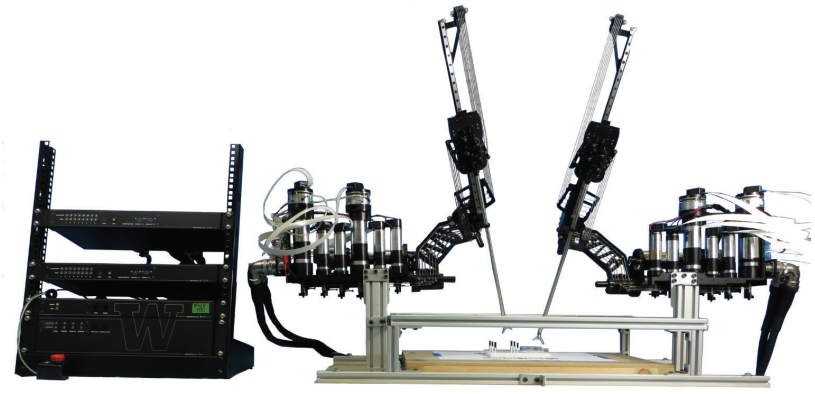
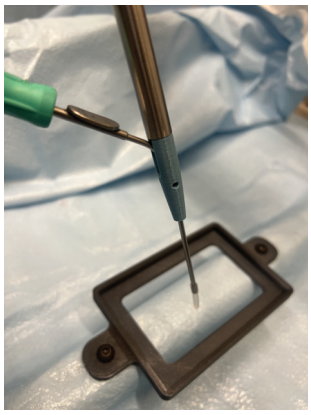


Figure 5.3: Overview of RAVEN II^{TM} Surgical System

5.3.4 Hardware

We use the RAVEN-II surgical robotics platform to demonstrate the setup of the framework. RAVEN-II [269] is an open platform designed for collaborative research in minimally invasive surgical procedures. Developed in 2012 by researchers at the University of Washington and the University of California - Santa Cruz, the RAVEN-II features two 3-degree-of-freedom (DoF) spherical positioning mechanisms. Surgical instruments with up to 4 DoF can be interchangeably attached, providing each arm with up to 7 degrees of freedom. The robot uses a cable-driven mechanism and was originally designed for teleoperative control, similar to the da Vinci surgical robot. When considering its application to autonomous tasks, it is necessary to address the limitations that affect positional precision. Limitations affecting positional precision include cable stretching, friction, mechanical backlash, and an imperfect kinematic model. Figure 5.3 presents an image of the RAVEN-II surgical robot and its controllers.

The RAVEN-II is used in conjunction with the Collaborative Robotics Toolkit (CRTK) [270], an interface specification that standardizes message formats for teleoperated and cooperatively controlled surgical robots. CRTK was inspired by the Canonical Robot Command Language(CRCL) [271] used similarly for industrial robots and automatic guided vehicles(AGV). CRTK was developed to enhance the compatibility of surgical robotics re-



(a) Modified Suction Instrument

(b) Suction Filter and vacuum

Figure 5.4: We use a RAVEN tool to accommodate a neurosurgical suction tool by incorporating a 3D-printed adapter. Blood from cavities will be removed using a vacuum pump and filter attached to the suction instrument.

search platforms with the da Vinci Research Kit(dVRK) [272]. dVRK utilizes components from first-generation da Vinci Surgical systems (Intuitive Surgical Inc.) along with software developed by Johns Hopkins University [273] and controllers created in collaboration with Johns Hopkins University and Worcester Polytechnic Institute. The goal of CRTK is to facilitate reproducible research by standardizing the specifications used across the different surgical robotics research platforms.

The peg transfer board is placed firmly on top of an 8020 using a 3D printed base (shown in Figure 5.4(a)). Considering a suction tool is symmetric about its roll axis, and the task requires only positional control of the tool tip, the movement of the RAVEN-II surgical robot is restricted to the first three DoFs. This allows the use of an operating mode called *Rick's tool* on the RAVEN, which constraints robot commands to be executed by the first three joints for a RAVEN tool. The tool, however, was initially designed as a grasper. To accommodate its use for a suction task, we designed a 3D printed adapter to attach the suction tube. Additionally, to ensure safe operation under unexpected contact at the tool

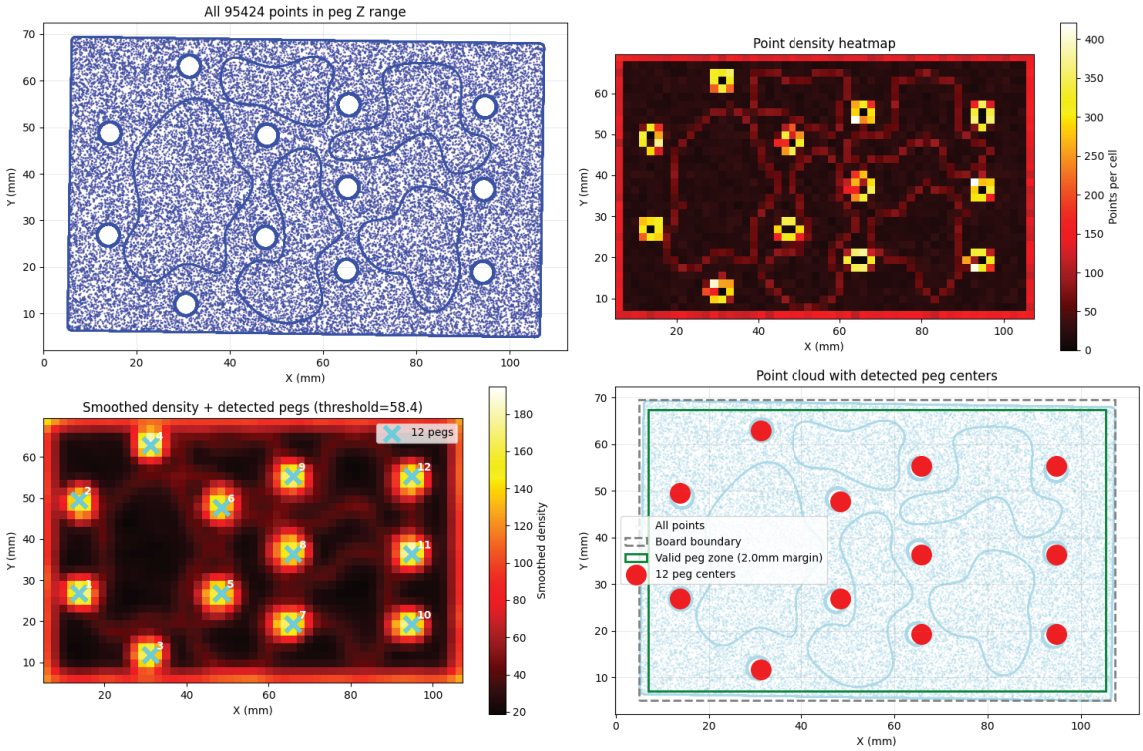


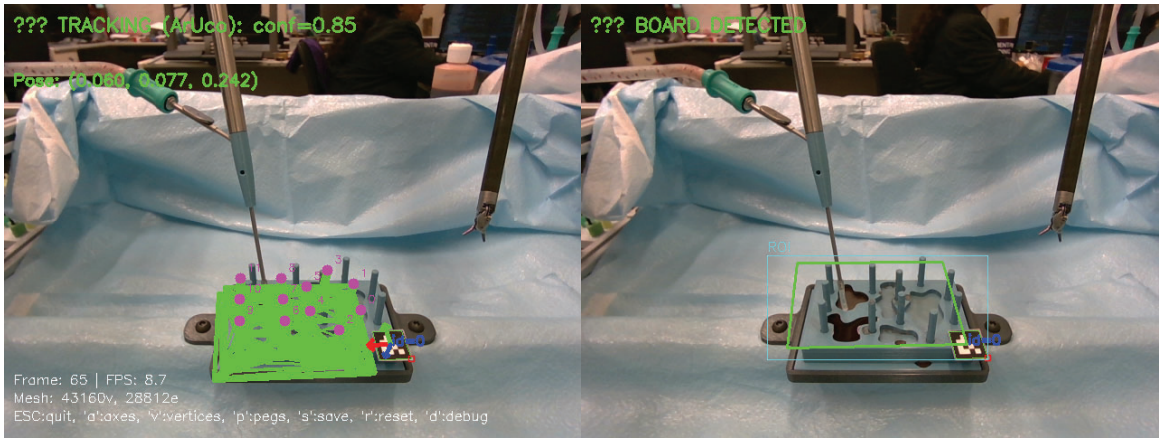
Figure 5.5: Location of Pegs on the task board, determined from point cloud data

tip, we attach a short piece of tubing to introduce compliance. The modified tool is attached to the right arm of the RAVEN-II surgical robot.

5.3.5 Collaborative Peg Transfer Task

The pegs on the task board present themselves as a static obstacle in the scene and are less likely to move while performing collision checks. To localize the pegs in the scene, we develop an algorithm that uses a density-based approach to transform 3D point cloud data into a 2D density map and estimate statistical peaks.

The algorithm for estimating peg locations is provided in Algorithm 2. This process assumes a known mesh of the task board is available in STL format. We sample points uniformly from the mesh to isolate regions with pegs by filtering based on the Z-coordinate values. Since the pegs are of a known height of 23.5mm and known X and Y positions, we



(a) Estimating board location (b) Board Region-of-interest (AruCo)

Figure 5.6: Board Localization in camera frame

constrain the search space to the location of the board. A 2D grid $\mathcal{G}$ over the XY-plane with a resolution of δ (in mm) discretizes the space. The value of δ is determined based on the size of the filtered point cloud. Following this, we estimate the density of points in the grid and apply a Gaussian Filter to reduce noise. We then apply peak detection and thresholding to identify the peaks. To reduce the likelihood of false positive peak detection, we constrain the peaks to occur within the known geometry of the board. To get the final set of peaks, we apply Density-Based Spatial Clustering of Application with Noise (DBSCAN) [274] to further refine multiple detections that are in close proximity to each other into one by estimating the centroid of nearby peaks.

We include the results of peak detection in Figure 5.5. The board localization is also performed using computer vision, by estimating the location using an Aruco marker and refining its pose via Iterative Closest Point [275]. A sample output is shown in Figure 5.6.

5.3.6 Motion Planning

The proposed framework comprises of two motion planning components that assess the effectiveness of a cooperative autonomy system. The first component evaluates the accuracy

of complete coverage planning for providing suction at a bleed point, which supports the methodology for testing Hypothesis 1 (H1). The second component implements dynamic motion planning to avoid interference with the surgeon's instrument during movement to the suction site and retraction upon task completion. Assistive autonomy is advantageous only if it does not disrupt the surgeon's actions or increase cognitive workload. Consequently, selecting suitable algorithms, such as those utilizing anytime planning and model predictive control, is essential for evaluating H2.

5.3.7 Metrics for evaluating collaborative autonomy

Hypothesis 1 (H1) evaluation relies on estimating the amount of residual blood in the surgical scene. The analysis to determine the efficiency in providing suction is performed by using image frames from a camera that captures the surgical site from a top-down view.

Surgical procedures are typically assessed through Objective Structured Assessment of Technical Skills (OSATS) [276] scored by surgical experts overseeing sessions. Despite OSATS being an effort towards objective skill assessment, there has been inter-rater variability that results in evaluations being subjective nevertheless. Kumar et al. [277] show a strong correlation between surgical skill analysis and motion metrics. Hence, to objectively measure the difference between performance in the two setups, we perform motion analysis using metrics associated with *instrument handling* and *time and motion* [278]. We define them as follows:

- *Average Smoothness* is the average change in acceleration of the suction instrument tip. We use Spectral Arc-Length as the smoothness measure [279].

- *Suction Path Length* is the total distance traveled by the suction instrument tip (2D & 3D)

- *Surgeon Path Length* is the total distance traveled by the surgeon's instrument tip (2D & 3D)

- *Average Velocity* of the suction instrument tip (2D & 3D)

- *Average Acceleration* of the suction instrument tip (2D & 3D)

- *Total Operative Time for Robot and Surgeon combined* is the total completion time for the combined experiment

- *Robot Operative Time* is the time for completion for the robot

- *Surgeon Operative Time* is the time for surgeon to complete the peg transfer task

- *Idle Time* is the total time for which the instrument is still. This used the information about instrument velocity and uses a pre-defined threshold to designate idleness. We start with 5mm as defined in [278] and will modify as appropriate for RAVEN II during the development process.

- *Economy of Volume* estimates the ratio of space traveled by the instrument to path length defined by

$$EoV = \frac{\sqrt[3]{[\max(X) - \min(X)] \cdot [\max(Y) - \min(Y)] \cdot [\max(Z) - \min(Z)]}}{pathlength} \tag{5.1}$$

- *Number of collisions* In each case of surgical suction, we calculate the number of collisions between the surgical instruments to provide suction.

Additionally, for Hypothesis 2 (H2), the results from a self-reported evaluation to measure task load - National Aeronautics and Space Administration Task Load Index (NASA TLX) [280] and Surgery-specific workload measure (SURG TLX) [281] - after users interact with the system will provide a comprehensive analysis. This accounts for mental demand, physical demand, temporal demand, performance, effort, frustration and total workload score. A paired t-test to evaluate the differences in performance metrics indicated above and workload metrics between the two conditions will provide statistical insights.

5.4 Conclusion

The authors propose a framework to benchmark cooperative autonomy for surgical robotics. The surgical system incorporates various components involving surgical scene understanding, motion planning, feedback control, and user studies.

5.5 Acknowledgements

The author would like to thank Jonathan Shu and Andrew Lewis from the University of Washington for their support in designing the 3D-printed parts described in this chapter.

Algorithm 2: Density-Based Peg Detection from Point Cloud

Require: STL mesh M, z-range $[z_{\min}, z_{\max}]$, grid resolution δ, threshold factor τ, Edge exclusion distance d_{edge}, max pegs max_{peak}, Proximity threshold ε

Ensure: Set of peg positions $\mathcal{P} = \{(x_i, y_i)\}$

$P \leftarrow \text{SamplePoints}(M, N)$ `// Sample N points from mesh`

$P_{\text{filtered}} \leftarrow \{p \in P : z_{\min} \leq p_z \leq z_{\max}\}$ $(x_{\min}, y_{\min}, x_{\max}, y_{\max}) \leftarrow \text{BoundingBox}(P_{\text{filtered}})$

Initialize density map $D \in \mathbb{R}^{\lceil (y_{\max}-y_{\min})/\delta \rceil \times \lceil (x_{\max}-x_{\min})/\delta \rceil}$ with zeros

for *each point* $p \in P_{filtered}$ **do**

  $i \leftarrow \lfloor (p_x - x_{\min})/\delta \rfloor$

  $j \leftarrow \lfloor (p_y - y_{\min})/\delta \rfloor$

  $D[i, j] \leftarrow D[i, j] + 1$

end

$\tilde{D} \leftarrow \text{GaussianFilter}(D, \sigma)$ $D_{\max} \leftarrow \text{MaximumFilter}(\tilde{D}, \text{window size})$

$\text{threshold} \leftarrow \tau \cdot \max(\tilde{D})$ $\text{Peaks} \leftarrow \emptyset$

for *each cell* (i, j) *in* $\tilde{D}$ **do**

  if $\tilde{D}[i, j] = D_{\max}[i, j]$ ***and*** $\tilde{D}[i, j] > threshold$ **then**

    $x \leftarrow x_{\min} + (i + 0.5) \cdot \delta$ $y \leftarrow y_{\min} + (j + 0.5) \cdot \delta$

    $\text{Peaks} \leftarrow \text{Peaks} \cup \{(x, y)\}$

  end

end

$\text{Peaks}_{\text{filtered}} \leftarrow \emptyset$

for *each peak* $(x, y) \in Peaks$ **do**

  if $x > x_{\min} + d_{edge}$ ***and*** $x < x_{\max} - d_{edge}$ ***and***

    $y > y_{\min} + d_{edge}$ ***and*** $y < y_{\max} - d_{edge}$ **then**

    $\text{Peaks}_{\text{filtered}} \leftarrow \text{Peaks}_{\text{filtered}} \cup \{(x, y)\}$

  end

end

if $|Peaks_{filtered}| > max_{peak}$ **then**

  $\text{Clusters} \leftarrow \text{DBSCAN}(\text{Peaks}_{\text{filtered}}, \varepsilon, \text{minPts} = 1)$

  $\mathcal{P} \leftarrow \emptyset$

  for *each cluster* $C \in Clusters$ **do**

    $\bar{x} \leftarrow \frac{1}{|C|} \sum_{(x,y) \in C} x$ $\bar{y} \leftarrow \frac{1}{|C|} \sum_{(x,y) \in C} y$

    $\mathcal{P} \leftarrow \mathcal{P} \cup \{(\bar{x}, \bar{y})\}$

  end

else

  $\mathcal{P} \leftarrow \text{Peaks}_{\text{filtered}}$

end

return $\mathcal{P}$

Chapter 6

SOCIAL SHAPING OF SURGICAL ROBOT DESIGN

Robotic surgery has been broadly advertised for its minimal invasion, reduced risk of infection, faster recovery, and lessened scarring to patients. With over 251,000 US deaths each year owing to preventable adverse events by clinicians [282, 283], robotics and autonomous surgery are marketed to surgeons as a potentially safer tool due to their increased dexterity and vision in hard-to-reach regions. Studies have found that approximately 70% of preventable medical errors are due to the breakdown of communication and collaboration between the operating room teams [178, 179, 284]. The strict hierarchical structures and asymmetrical power dynamics between members of the medical team can deter them from reporting these medical errors [179, 285]. This highlights the importance of the "social" being co-located with the "technical" while working with safety-critical systems. In the context of surgical robotics in the operating room, social means the people involved and technical refers to the structure and processes that shape interactions. While the robot-assisted surgery scholarly literature emphasizes the increased technical performance afforded by a robot [286, 287], it does not fully consider its impact on the working environment for the *actors* involved.

Researchers have discussed concerns over job losses in the medical sector with the introduction of intelligent, autonomous robots in the operating room [288] and the ethics of including robotic caregivers [289, 290]. However, analysis of the changes to the roles of the medical staff introduced by the surgical robot has been limited [39, 291]. The surgeon, who is now hunched over a display monitor to view the input from an endoscopic camera, is no longer located in close proximity to their patient or their team. The room is arranged around the patient, who is located towards the center and placed adjacent to a massive robot base with four surgical arms. The surgical team is spread around the room, prepared to step in as required. Fosch-Villaronga et al. [292] explore the shifting relationship of

the different stakeholders as increasing levels of autonomy is introduced into the operating room. The authors argue that potential benefits to surgical automation cannot be realized unless more clarity on responsibilities between the different actors is achieved.

Surgical robots and the paradigm shift they represent raise a host of ontological, ethical, and technological questions. We argue that the consequences of the use of surgical robots in operating rooms calls for a sociological examination of their creation. By using Social Construction of Technology (SCOT) [293], this aim will analyze literature around the actors involved with designing technical systems, specifically about surgical robotics design, to obtain a more robust understanding of the dynamics between technological and social elements that shaped its current development and use. Here, the surgical robot is not viewed as a tool but rather as a socio-technical relation that integrates the mechanical equipment and working tools, the operators, the patient and other stakeholders. We present an initial summary of the social origins of surgical robotics below.

6.1 Social Origins of Surgical Robotics

While analyzing the Human-Robot Interaction design for a surgical robot, it is important to understand the historical context of the system's development. We will focus on the origins of teleoperated robots for surgery due to their large-scale adoption and its evolving status as the standard of care for minimally invasive surgery [294]. While a number of publications focus on the chronology of robots in the field [295,296], they do not engage with the design implications of surgical robots due to funding agencies.

The field of intelligent robotics has made headway from being a concept to being actively integrated into our society over the last century. The idea of robots performing exceptionally complex tasks such as surgery on patients seemed like science fiction, and it was for a long time. In 1925, Dr. Hugo Gernsback, most famous for his science fiction magazines[1], published what is possibly the first concept of a teleoperated device for medical applications in *Science and Invention* [297]. Gernsback's 'teledactyl' illustration (shown in Figure 6.1(a))

[1]We would like to acknowledge Gernsback's contribution to science fiction while not negating his venality, corruption, and disregard for the financial rights of authors. Interested readers can find more information in Resnick, Mike; Malzberg, Barry (Dec 2009 – Jan 2010). "Resnick and Malzberg Dialogues XXXXVI: The Prozines (Part 1)". The SFWA Bulletin. 43 (5): 27–28.

Figure 6.1: Teleoperated robotic surgery concept: (a) from Hugo Gernsback's illustration in 1925, (b) proposed by NASA Advanced Concepts and Missions Division in 1972

envisioned the use of robotic arms and wireless technology, where a doctor and his team would provide remote consultation for a patient. The patient is depicted to have a human assistant present in the room to help as required. To quote the concept as described in the publication, *"Here we see the doctor of the future at work, feeling the distant patient's arm. Every move that the doctor makes with the controls is duplicated by radio at a distance"*. Teleoperated systems allow users to use their expertise from a distance for scenarios that are either inaccessible or dangerous for human operators to be present in.

The first known research publication on the medical application of robotics came nearly 50 years after Gernsback's idea. In 1972, a research group from the National Aeronautics and Space Administration (NASA) Advanced Concepts and Missions Division provided a conceptual design of a robotic teleoperation system[2] and the use of dedicated medical-purpose communication satellites [298]. Telepresence offers the potential for medical teams to treat their patients even if they are geographically separated. NASA was interested in using interactive telepresence [299] to provide surgical care for astronauts by interfacing via

[2]The use of teleoperation in robots has been referred to as a "master-slave" setup in the literature. This community is looking to use alternate terminology instead (Link: https://www.eetimes.com/its-time-for-ieee-to-retire-master-slave/)

remotely operated robots controlled by expert surgeons located on Earth. The advancements in telecommunication technologies and cameras, combined with the progress made in space programs and robotics for handling nuclear material, provided the fundamental characteristics of the remote system design:

> Use of Manipulators for Remote Medical Emergency Care: Recent advances in the technology of human extension by remote manipulation, concepts derived largely from the U.S. space program, make possible professional medical care, including surgery, in situ at the emergency site and remote from an environment specializing in acute care (such as a hospital emergency facility or Clinic). Figure 2 (Figure 6.1(b) in this chapter) illustrates such a remote surgical system. A proposal, currently under consideration, suggests the use of human extension through remote manipulation to upgrade the quality of medical care at both remote and accessible areas and to make the most effective use of scarce skilled medical practitioners throughout this nation and the world for the relief of human suffering. [298]

Despite the early conversations around developing these systems, the field of Robot-Assisted Surgery (RAS) was only propelled by private-public partnerships some 15 years later [300, 301]. Many of the robotic systems developed around this period were primarily used for tool holding/positioning. For example, the Automated Endoscopic System for Optimal Positioning (AESOP) from Computer Motion (Goleta, CA) was used to hold an endoscopic camera during the procedure [302].

Research groups started developing and testing the viability of building surgical robots due to increased interest and availability of funding in the late 1980s. Experiments were performed to verify the use of remote healthcare paradigms. A general-purpose 1986 telepresence system using a Head Mounted Display (HMD) and DataGloves [300], for visualization and instrument control respectively, provided a method for virtual interactions with digital displays. This was repurposed by US Army Colonel (COL) Richard Satava, MD, and Phil Green, Ph.D., where they replaced the DataGloves on the user side with surgical instruments and used a stereoscopic monitor instead of the HMD [303]. This became the first

"telepresence surgery system". Following a few different iterations, the first prototype to impact remote patients was the Green Telepresence System, assembled in 1991 at Stanford Research International(SRI) [304,305]. The Green Telepresence System contributed to the current surgical robot design we see in procedures. Minimally invasive laparoscopic surgery, with manual instruments and a wall-mounted monitor, was gaining traction around this period due to reduced scarring for patients and quicker recovery time. While the telepresence system was conceived for open surgery, the momentum from the surgical community led to the transition of its use for laparoscopic procedures [301]. The leading researchers [306,307] in surgical robotics suggested laparoscopic interventions as "transitional" technology in the history of surgical evolution. In subsequent years, telepresence surgery research groups conducted experiments with systems to perform intercontinental surgery [308–310], underwater surgery to analyze latency effects [311], and operated on rats in micro- and zero gravity [312,313].

Following NASA's concept of telesurgery, the Department of Defense (DoD) was interested in providing care to battlefield injuries, and telepresence surgery offered a possible solution by removing distance as a factor in providing immediate and intensive treatment. The use of teleoperated robots was also intended to address the issues of restricted accessibility and limited support personnel. Providing care at the site of injury and/or during transport was viewed as more favorable than evacuating the injured to the closest battalion aid station. The robots were envisioned not to be a tool but rather an information system. It would combine different capabilities such as telecommunications, remote imaging, computer-enhanced manipulation, image acquisition, and guidance [314].

Based on the experiences at SRI and NASA, in 1999 the Defense Advanced Research Projects Agency (DARPA) and US Army Medical Research and Materiel Command (MRMC)/ Telemedicine and Advanced Technology Research Center (TATRC) aimed to develop the Digital Human Robotic Casualty Treatment and Evacuation Vision with robotic systems. One of the resulting projects was a deployable trauma care system called the 'Trauma Pod'. This system, according to original plans, would allow combat surgeons to perform life-saving operations on wounded soldiers from a distance by 2025 [315]:

> The main goal was to "enhance battlefield casualty care by developing autonomous and semi-autonomous mobile platforms through the integration of tele-robotic and robotic medical systems. The initial phase has successfully automated functions typically performed by the scrub nurse and circulating nurse... The next phase of the program will develop methods for autonomous airway control and intravenous access... Finally, these systems will be miniaturized and incorporated into a tactical platform capable of operating in a battlefield or mass casualty environment." [316]

In 2005, DARPA Defense Sciences Office selected and funded SRI to lead the Trauma Pod project. It achieved its stated goal of automated tool change on the da Vinci within 10 seconds. While the project has not been deployed on the battlefield, prototypes have shown initial success [317]. The U.S. Army continues to be interested in technologies for remote care delivery on the battlefield through work at TARTC.

Though initial experiments and prototypes were aimed at long-distance telesurgery, only a few cases have performed long-distance procedures following the commercialization of surgical robots [318–321]. Some of the key challenges with remotely operated robots for surgery has been the communication lag time, limited bandwidth, latency, and loss of signal. Despite their shortcomings for long-distance procedures, these projects created the basis for the most commercially viable and successful surgical robot, the da Vinci from Intuitive Surgical Inc. (Sunnyvale, CA) [21] and the discontinued ZEUS Robotic Surgical System. These robots provide a reliable assistive teleoperation platform for surgeons to operate on patients by offering reduced hand tremors, increased range of motion, and virtual magnification of the surgical area during procedures.

With the field initially being catapulted by research interests in space and military care, the robotic systems were designed for remote care and focused on human-robot interaction processes centered around the surgeon and the patient. When these devices were commercialized, they were integrated into existing operating rooms where medical teams were involved, comprising a number of other actors such as the assistants to the surgeon, scrub nurse, circulating nurse, technicians, and anesthesiologists among others. Due to the per-

ceived flexibility in design, surgical robots have been used for procedures they were not intended for [322]. In subsequent sections of this chapter, we will describe the effect of integrating teleoperated robotic systems into present-day operating rooms, and identify the relationships between robot design and the impact on the surgical team.

6.1.1 Human-Robot Interaction

The rapidly growing adoption of surgical robots in the operating room [19] promises myriad benefits [286, 323]. While the advantages to patient outcomes (reduced scarring and pain, shorter hospitalization [36]) and surgeons (enhanced dexterity and ergonomics, greater precision [287]) are regularly discussed, studying the effect of introducing a complex interactive system into the operating room team setting has been limited [291, 324]. Though the exact benefits of robotic surgery over minimally invasive surgery for patients have not been clearly established, with studies showing negligible benefits due to increased costs [325,326], hospitals are interested in increasing their use of robot-assisted procedures due to benefits offered to surgeons [19]. To successfully integrate robotic surgery into the operating room, the personnel involved have to adapt individually and as a team. Important factors include a dedicated robotic team [327] and additional staff [328]. Surgical work is highly collaborative, with surgical members relying on the skills of the entire team [329, 330]. Design choices in RAS could potentially result in unintended consequences for a team setting which may not be obvious to the system developer. While the interpretive flexibility [293] of the da Vinci system has extended its application beyond laparoscopic surgery, the use of the artifact in ways that could lead to potential patient safety hazards should not be discounted [331].

6.1.2 Teamwork in Operating Rooms

A 2022 scholarship analyzing medical errors determined that approximately 70% of all medical errors are a result of communication breakdown within the healthcare team! [332]. Research by patient safety experts has identified communication and teamwork to be critical in preventing and mitigating medical errors [333]. Considering Medical Errors were determined to be the third leading cause of deaths in the United States [282] and Pre-

ventable Medical Errors (PME) account for over 400,000 patient deaths according to a recent study [179], it is important for surgical robotics researchers to consider social factors in their design approach. Here, medical error is defined as "... an unintended act (either of omission or commission) or one that does not achieve its intended outcome, the failure of a planned action to be completed as intended (an error of execution), the use of a wrong plan to achieve an aim (an error of planning), or a deviation from the process of care that may or may not cause harm to the patient."

More recently, a study by Rogers et al. [334] analyzed surgical errors that led to malpractice claims with communication breakdown contributing to 24% of the claims. A different research group deduced that team communication challenges led to 43% of adverse events [335]. While the studies were conducted on different conditions and outcomes, it is evident that team dynamics are crucial while working in a safety-critical environment. Even if the communication failures do not result in what is professionally designated as "adverse" [336], it could still affect the team's ability to work effectively under uncertainty [337].

With the doctor being located farther away from the patient in an operating room, there are different factors that shape how the team interaction is changing. We have categorized the literature into overarching themes that can be addressed using better design processes - team communication, spatial configuration, and situational awareness.

Verbal and Non-verbal Communication:A surgeon requires the help of team members to effectively perform procedures [338, 339] on the patient-side while owing to the physical separation. Studies comparing communication in traditional laparoscopic surgery to robotic surgery determined that the latter required significantly greater oral communication from the surgeons to compensate for the distance from the patient [340–342]. The team finds it challenging to hear the directions from the surgeon due to the immersive nature of the surgeon's console [343]. The surgeon is required to bend down to get a continuous visual stream of the surgical site and the positioning results in a muffled voice. Additionally, this requires the team to pay much closer attention to verbal communication from the surgeon, which could otherwise result in potential PMEs. An observational study from Griffith University [344] found that communication failure could be a predictor of deviations to the estimated time of a surgical procedure.

Randell et al. [324] conducted a multi-site interview study with operating room teams to analyze the impact of surgical robots on decision making and the contexts in which these impacts are likely to occur. The realist evaluation approach used by the group involves "gathering data in order to explain how different contexts trigger particular changes in the reasoning and responses of users ('mechanisms') which, in turn, give rise to a particular pattern of outcomes" (Pg. 6). Participants of the study were presented with ideas from the literature and compared them with their personal experience of using robot-assisted surgical platforms. The findings stressed the importance of a positive relationship between the surgeon and the operating room team, without which there could be complications or increased distractions. This is bolstered by [345], which studies communication between the chief surgeon and surgical assistant performing robot-assisted procedures. The level of interdependency in tasks performed by the surgeon and their assistant is reduced, which may alleviate some of the challenges with team communication between them [346]. Finally, trainees play a more passive role in robotic surgery, spending a majority of time observing the surgeon as opposed to actively participating [347].

In addition to verbal cues, the increased separation of the team also limits the non-verbal cues surgeons can use for communication including gestures, eye gaze, and body orientation [348, 349]. This also poses a challenge for a surgeon's ability to use non-verbal cues when verbal cues are insufficient.

Spatial Configuration: The large size of today's surgical robot reconfigures how the team operates around the patient within the OR. According to [350], the spatial configuration was determined to be a "performance-shaping" factor for surgery. The reconfigured setup for robotic surgery requires the surgeon to communicate surgical tool requirements more explicitly. These directional cues can result in confusion, additional time, and possible patient injuries [345]. The ability of the team to anticipate decisions is challenging in these reconfigured workspaces. For example, the anticipatory behavior that is characteristic of scrub nurses for team coordination does not hold as well during robotic procedures. Additionally, changing to open surgery or performing emergency interventions becomes challenging due to the physical constraints presented by the robot, which occupies a large volume of space within the operating room and has to be carefully navigated to avoid disruptions [338].

A study by [291] reports findings from two years of ethnographic observation and video analysis through fieldwork performed at two hospitals. The primary focus is to address "How do teleoperated surgical robots reconfigure the spatial arrangement of surgical teams in the operating room?" and "How does this spatial reconfiguration affect and remediate the work of broader surgical teams, including in their material, experiential, and affective dimensions?". The authors conclude that adding a robot reconfigures the spatial arrangement of surgical teams in three ways - (1) placing the surgeon at a distance from the team, (2) scattering team members around the patient and (3) introducing a visual barrier between team members.

Situational awareness: Situation awareness is defined as the "perception of elements of the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" [351]. [352] and [353] found that increased situational awareness in surgeons could result in fewer errors during surgery. The positioning of members from the clinical team affords them with different information depending on their role within the team to inform decision-making. The surgeon is now dependent on the team for information due to physical and contextual isolation [354].

Open surgery has shown better shared awareness within the team, resulting in an uninterrupted flow of information [329]. The study by Randall et al. [324] highlighted the importance of a team in ensuring the surgeon is situationally aware by communicating the state of not only the patient but also the robot. The team has to understand decisions from the context of the surgeon, which is impeded by the immersive viewers provided primarily for surgeons. Though there are multiple displays present in the operating room displaying the movement of instruments, the informational resources are not equally available for everyone in the team. This should be viewed as a two-way dependency and hence studies should not only focus on making the surgeon situationally aware through design decisions, but also propose ways to increase situational awareness for the entire team. To deal with complications, the team should have a shared situational awareness of the procedure [338].

6.1.3 Machine Learning and Human-Robot Interaction in Teams

Surgical expertise includes the decision-making process [355] within the operating room. Previous experiences with respect to surgical scenarios allow the surgeon and the team to use a shared mental model for providing optimal patient outcomes. Selection and use of such a model includes focusing on relevant cues, providing expectations, and identifying goals and tactics/techniques appropriate for the scenario [356]. Research studies have strongly demonstrated that the technology and physical proximity of team members play a key role in collective decision making [357–359].

Over the last few years, there has been an increased interest in integrating machine learning (more generally referred to as Artificial Intelligence or AI) with surgical robots and the operating room. Collecting large volumes of data has never been easier and researchers are mining video, audio and instrument motion data to gather relevant patterns to propel augmenting or automating parts of a surgical procedure [29, 360], provide increased contextual awareness for surgeons [361, 362], or for evaluating the skill level of surgical procedures [363].

Human-AI partnerships require a human operator to develop insights into the decisions made by the "technical agent" (or AI), by forming a mental model for shared decision-making. Bansal et al. [364] posit that updates from developers to the agent could affect this shared mental model, even if the performance accuracy of the AI increases. They hypothesize that human-AI teams become susceptible to failures when there are discrepancies introduced by a system that does not account for the human users' mental model developed from previous interactions. By performing empirical studies on three different machine learning applications, including one in a clinical setting, the authors show that developing a synergetic mental model improves performance of the human-AI team. However, model updates only focused on AI performance may degrade net team performance. Relating this to our application within the operating room, updates sent to a surgical robot's machine learning model may increase the overall performance of the device, but the implications of these model updates on the surgeon and the broader team's performance, and especially patient outcomes, need to be studied. As surgical procedures start using automated techniques, it is important to include the robot as an independent actor as opposed to a tool.

Drawing from observations in [365], we hypothesize that the clinical staff and operating room technicians will play a central role in the success of autonomous robotic agents within the operating room.

Deutsch et al. suggest that the immersive nature of current robotic systems in the operating room reduces distractions for surgeons [366]. This could be a benefit for surgeons to avoid case-irrelevant discussions, which could result in additional intraoperative stress and increase a surgeon's mental fatigue [367]. However, a different study found that complex technological systems could result in more frequent distractions due to the equipment and distractions associated with the work environment [368], negating one of the proposed advantages for initial users of the robot.

6.2 Discussion

The widespread use of surgical robotics as we know it today was built on top of research interest and funding from space and military agencies. The interest from NASA and the Department of Defense combined with private-public partnerships led to a number of startups being founded to work on robot-assisted surgical technology. Despite the large number of robots prototyped for medical applications in research labs, very few make it to the market. Most failed due to the strict policies implemented by organizations like the Food & Drug Administration (FDA) which closely monitor the safety and effectiveness of medical systems being deployed [369].

Beyond the approval and commercialization of these robots, understanding how the technology is diffused provides an insight into the role of early adopters. Companegi et. al make [370] use of quantitative and qualitative methods to study how surgical robotics spread in Italy between 1999 and 2010. The authors analyzed the "motivational drivers" that led to the adoption of surgical robotics and the mechanisms involved in the dissemination of "implementation experiences". Their work describes how central actors versus peripheral actors frame a new piece of technology and how they communicate about the technology to the public. This social group of peripheral adopters, constituting themselves as "exemplary users", plays a primary role in the shaping of technology and its eventual adoption by users. The information gathered from interviews with adopters versus non-adopters shows

the influence of restrictions placed by regional governments and the presence of budgetary restrictions, among other factors. They determined that some of the main events that aided diffusion of Robot-Assisted Surgery was live surgical demonstrations with robots at national conferences, starting conferences centered around robotic surgery, and the creation of societies that would work on communication and dissemination of knowledge. This highlights the role of institutional infrastructures that contribute to determining hierarchies and consolidating social position within a community.

Our review of surgical robots, initially designed for the military and space and then integrated into operating rooms, determined that physical distance between the members of the team and the robot has changed how human-human and human-robot interactions occur in the OR. These challenges in team settings include developing shared mental models [371, 372], limited "perception" despite fusing information from multiple sensors [371, 372], and reduced situational awareness [371, 373, 374], absence of nonverbal cues [375], and reduced influence of members beyond the central actors [376]. These challenges persist even though the team is still collocated inside the operating room, in contrast to proposed long distance teleoperated robots.

The challenges of coordinating a surgical team with advanced robotics, reviewed in this paper, present a unique set of opportunities to improve surgical robot design by supporting breaking down strict hierarchical structures within clinical teams. The procedures require the team to be more collaborative and vocal about any surgical and/or technical concerns using verbal communication. This shifts the power dynamics and empower individual members [291]. Deutsch et al. suggest that the immersive nature of current robotic systems in the operating room reduces distractions for surgeons [366]. This could be a benefit for surgeons to avoid case-irrelevant discussions, which could result in additional intraoperative stress and increase a surgeon's mental fatigue [367]. However, a different study found that complex technological systems could result in more frequent distractions due to the equipment and distractions associated with the work environment [368], negating one of the proposed advantages for initial users of the robot. Furthermore, the lack of tactile feedback continues to be a limitation that researchers need to consider in their design process [377]. This would present increased situational awareness to the surgeon, reducing some of the

stress within a surgical team.

The side effects of viewing society (for example, as its needs are expressed through the regulatory process, or as financial impacts are exposed through business practices) as an externality in the design process are disadvantageous, especially in a safety-critical system where we should aim to reduce cognitive load. The social context of the system's use and its implications on users are only being considered after the entire development process, when the surgical robots are ready to be deployed and little can be changed. Robot design is more than the simple production of technological artifacts and includes "narratives about social order, human behavior and psychology, and common norms, beliefs, desires and expectations" [378], also known as technoscientific imaginaries. Therefore making conscious and/or unconscious decisions while developing technology can "open certain social options and close others" [379].

With new technological advancements, such as Augmented Reality (AR) for surgery [236], that aspire to integrate into the operating room, developers should make proactive design decisions for a team setting. While it is easier to partake in the "single point" notion of interaction between humans and robots, designing with a sociological perspective will make the design more relevant and useful [380]. The future of HRI will be shaped by the different levels of autonomy in surgical robotics. The authors in [36] propose six different levels of autonomy that utilize varying magnitudes of user involvement. The authors describe the regulatory challenges involved with each of these levels, including the role played by the FDA in the approval process of robotic technology for healthcare. The FDA, and comparable regulatory agencies worldwide, will shape when and who gets to experience medical technology. Designing robotic interfaces by accounting for social and team dynamics factors may thus allow to gently nudge teams towards effective collaboration practices in cognitive and affective terms [291].

6.3 Conclusion

In this chapter, we have reviewed the literature on different societal factors involved in shaping the design of the most widely used surgical robot, challenges involved with the adoption of the technological artifact, attributes of teamwork that have been modified due

to the introduction of robots in ORs, and summarized how machine learning and shared mental models will play a key role in this evolving standard of care.

Chapter 7

CONCLUSION AND FUTURE DIRECTIONS

The goal of my research is to develop tools for surgical scene understanding to enable collaborative human-robot interaction for surgical assistance.

In Phase 1, my research focuses on developing a surgical instrument segmentation framework to identify instruments located in an image at a pixel-level. We demonstrate the integration of transfer learning, data augmentation and a parameter-efficient backbone with a convolutional neural network developed based on a U-net type encoder-decoder architecture to improve the performance for both endoscopic robotic instrument segmentation and microneurosurgical instrument segmentation. We also released the NeuroID dataset for public access through this work.

In Phase 2, we implemented and evaluated a framework using epistemic uncertainty for active learning and identifying data points for annotation in surgical instrument segmentation tasks. Using this approach, we were able to reduce the amount of annotated data required to achieve desired performance. This also helps increase the efficiency of samples acquired for labeling. Our work also studies the different visualization tools that aid in understanding where the model currently fails using uncertainty estimation and could improve the understanding of the model for users. This analysis benefits other applications that utilize segmentation models such as biomedical image segmentation, self-driving vehicles and general robotic tasks such as manipulation.

While improved performance is achieved with fewer data points using active learning, a large number of annotated data points are still needed to transfer performance to new surgical settings and conditions. To overcome these shortcomings, we tap into the benefits of pre-training by using a foundation model. The objective was to use pre-training to use low-dimensional representations of a large volume of data for finetuning to surgical data and solve downstream surgical tasks of interest. We proposed an improved model to perform

reference-based surgical instrument segmentation by utilizing the annotation data available in benchmark datasets. We also provided an analysis on potential pitfalls of using AI for assistive data annotation and the implication of noise in the labels.

In Chapter 05, we describe a benchmarking framework that incorporates surgical scene understanding components to enable robotic assistive suction while working cooperatively a surgeon. In future work, the performance of the autonomous system will be compared against that of human assistance via teleoperation of the same robot assistant to assess the system's benefits in terms of surgical performance and workload demands. Although beyond the scope of my PhD, I believe that the derived methodology could be eventually extended to work with a context-aware system, where the robot has knowledge about the current phase within a surgery. This would help transform the system from Level 2 autonomy to level 3 autonomy.

In order to realize synergistic human-robot interaction (HRI) both at the surgeon-robot level and surgical team-robot level, it is important to understand the current landscape of HRI in surgery. The goal of work done in Chapter 06 was to survey the current surgical robotics literature and outline design factors to consider while building robots for autonomous functions in the operating room.

7.1 Future Directions

- Based on experience from Chapter 3 and 4, a potential direction of interest is in exploring how we perform data annotation in the age of foundation models. Considering the multifold improvements in performance for many downstream tasks from using pretrained models trained on internet-scale data, and the ability to generate labels while using point or text prompts, the efficiency or using select regions or phrases for prompting could be studied.

- To move towards higher levels of autonomy, the system proposed in Chapter 5 would benefit from integrating contextual information about the surgery [381–383] through techniques like surgical workflow analysis [384] or activity recognition for emergency response proposed in [385].

- A long-term limitation for developing surgical robot learning at scale had been the availability of a reliable physics-based simulator, addressed recently by the release of ORBIT-Surgical from Nvidia [386]. It would be interesting to use this state-of-the-art physics simulator to study safety concepts for collaborative assistance in suturing through a data-driven approach.

- Inferring the intention of the user by modeling their behavior could yield interesting insights into performing cooperative autonomy.

BIBLIOGRAPHY

[1] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International journal of robotics research*, 37(4-5):405–420, 2018.

[2] Harish Ravichandar, Athanasios S Polydoros, Sonia Chernova, and Aude Billard. Recent advances in robot learning from demonstration. *Annual review of control, robotics, and autonomous systems*, 3:297–330, 2020.

[3] Nicholas Roy, Geoffrey Gordon, and Sebastian Thrun. Planning under uncertainty for reliable health care robotics. In *Field and Service Robotics: Recent Advances in Reserch and Applications*, pages 417–426. Springer, 2006.

[4] Brady Zhou, Philipp Krähenbühl, and Vladlen Koltun. Does computer vision matter for action? *Science Robotics*, 4(30):eaaw6661, 2019.

[5] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[6] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[8] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):5, 2021.

[9] Giuseppe Turchetti, Ilaria Palla, Francesca Pierotti, and Alfred Cuschieri. Economic evaluation of da vinci-assisted robotic surgery: a systematic review. *Surgical endoscopy*, 26:598–606, 2012.

[10] Paolo Zaffino, Sara Moccia, Elena De Momi, and Maria Francesca Spadea. A review on advances in intra-operative imaging for surgery and therapy: imagining the operating room of the future. *Annals of Biomedical Engineering*, 48:2171–2191, 2020.

[11] Kyle H Sheetz, Jake Claflin, and Justin B Dimick. Trends in the adoption of robotic surgery for common surgical procedures. *JAMA network open*, 3(1):e1918911–e1918911, 2020.

[12] Michael R Mathis, Milo C Engoren, Hyeon Joo, Michael D Maile, Keith Aaronson, Michael Burns, Michael Sjoding, Nicholas J Douville, Allison M Janda, Yaokun Hu, et al. Early detection of heart failure with reduced ejection fraction using perioperative data among non-cardiac surgical patients: A machine learning approach. *Anesthesia and analgesia*, 130(5):1188, 2020.

[13] Haishuai Wang, Zhicheng Cui, Yixin Chen, Michael Avidan, Arbi Ben Abdallah, and Alexander Kronzer. Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(6):1968–1978, 2018.

[14] Russell H Taylor, Arianna Menciassi, Gabor Fichtinger, Paolo Fiorini, and Paolo Dario. Medical robotics and computer-integrated surgery. In *Springer handbook of robotics*, pages 1657–1684. Springer, 2016.

[15] N Hubert, M Gilles, K Desbrosses, JP Meyer, Jacques Felblinger, and J Hubert. Ergonomic assessment of the surgeon's physical workload during standard and robotic assisted laparoscopic procedures. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 9(2):142–147, 2013.

[16] Ian Jun Yan Wee, Li-Jen Kuo, and James Chi-Yong Ngu. A systematic review of the true benefit of robotic surgery: Ergonomics. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 16(4):e2113, 2020.

[17] RH van‘ van der Schatte Olivier, CDP van ‘t Hullenaar, JP Ruurda, and IAMJ Broeders. Ergonomics, user comfort, and performance in standard and robot-assisted laparoscopic surgery. *Surgical endoscopy*, 23:1365–1371, 2009.

[18] Thomas M Ward, Pietro Mascagni, Yutong Ban, Guy Rosman, Nicolas Padoy, Ozanan Meireles, and Daniel A Hashimoto. Computer vision in surgery. *Surgery*, 169(5):1253–1256, 2021.

[19] R. Elek, T. Dániel Nagy, D. Á. Nagy, G. Kronreif, I. J. Rudas, and T. Haidegger. Recent trends in automating robotic surgery. In *2016 IEEE 20th Jubilee International Conference on Intelligent Engineering Systems (INES)*, pages 27–32, June 2016.

[20] Simon DiMaio, Mike Hanuschik, and Usha Kreaden. The da vinci surgical system. *Surgical robotics: systems applications and visions*, pages 199–217, 2011.

[21] Gary S Guthart and J Kenneth Salisbury. The intuitive/sup tm/telesurgery system: overview and application. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, volume 1, pages 618–621. IEEE, 2000.

[22] Intuitive Surgical Inc. Annual report 2022, 2022.

[23] Aleks Attanasio, Bruno Scaglioni, Elena De Momi, Paolo Fiorini, and Pietro Valdastri. Autonomy in surgical robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:651–679, 2021.

[24] Simon A Menaker, Sumedh S Shah, Brian M Snelling, Samir Sur, Robert M Starke, and Eric C Peterson. Current applications and future perspectives of robotics in cerebrovascular and endovascular neurosurgery. *Journal of NeuroInterventional Surgery*, 10(1):78–82, 2018.

[25] Sharib Ali. Where do we stand in ai for endoscopic image analysis? deciphering gaps and future directions. *npj Digital Medicine*, 5(1):184, 2022.

[26] George P. Moustris, Andreas I. Mantelos, and Costas S. Tzafestas. Shared control for motion compensation in robotic beating heart surgery. In *2013 IEEE International Conference on Robotics and Automation*, pages 5819–5824, 2013.

[27] Kirsten E Kaplan, Kirk A Nichols, and Allison M Okamura. Toward human-robot collaboration in surgery: Performance assessment of human and robotic agents in an inclusion segmentation task. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 723–729. IEEE, 2016.

[28] Kamran Shamaei, Yuhang Che, Adithyavairavan Murali, Siddarth Sen, Sachin Patil, Ken Goldberg, and Allison M Okamura. A paced shared-control teleoperated architecture for supervised automation of multilateral surgical tasks. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1434–1439. IEEE, 2015.

[29] Adithyavairavan Murali, Siddarth Sen, Ben Kehoe, Animesh Garg, Seth McFarland, Sachin Patil, W Douglas Boyd, Susan Lim, Pieter Abbeel, and Ken Goldberg. Learning by observation for surgical subtasks: Multilateral cutting of 3d viscoelastic and 2d orthotropic tissue phantoms. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1202–1209. IEEE, 2015.

[30] Juan Antonio Barragan, Daniela Chanci, Denny Yu, and Juan P Wachs. Sachets: Semi-autonomous cognitive hybrid emergency teleoperated suction. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 1243–1248. IEEE, 2021.

[31] François Chadebecq, Laurence B Lovat, and Danail Stoyanov. Artificial intelligence and automation in endoscopy and surgery. *Nature Reviews Gastroenterology & Hepatology*, 20(3):171–182, 2023.

[32] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[33] Max Allan, Alexey Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, Luis Herrera, Wenqi Li, Vladimir I. Iglovikov, Huoling Luo, Jian Yang, Danail Stoyanov, Lena Maier-Hein, Stefanie Speidel, and Mahdi Azizian. 2017 robotic instrument segmentation challenge. *CoRR*, abs/1902.06426, 2019.

[34] Burr Settles. Active learning literature survey. 2009.

[35] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36. JMLR Workshop and Conference Proceedings, 2012.

[36] Guang-Zhong Yang, James Cambias, Kevin Cleary, Eric Daimler, James Drake, Pierre E Dupont, Nobuhiko Hata, Peter Kazanzides, Sylvain Martel, Rajni V Patel, et al. Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy, 2017.

[37] Paolo Fiorini, Ken Y Goldberg, Yunhui Liu, and Russell H Taylor. Concepts and trends in autonomy for robot-assisted surgery. *Proceedings of the IEEE*, 110(7):993–1011, 2022.

[38] Hang Su, Andrea Mariani, Salih Ertug Ovur, Arianna Menciassi, Giancarlo Ferrigno, and Elena De Momi. Toward teaching by demonstration for robot-assisted minimally invasive surgery. *IEEE Transactions on Automation Science and Engineering*, 18(2):484–494, 2021.

[39] Rebecca Randell, Stephanie Honey, Natasha Alvarado, Alan Pearman, Joanne Greenhalgh, Andrew Long, Peter Gardner, Arron Gill, David Jayne, and Dawn Dowding. Embedding robotic surgery into routine practice and impacts on communication and decision making: a review of the experience of surgical teams. *Cognition, Technology & Work*, 18(2):423–437, 2016.

[40] Manjunath Siddaiah-Subramanya, Kor Woi Tiang, and Masimba Nyandowe. A new era of minimally invasive surgery: progress and development of major technical innovations in general surgery over the last decade. *The Surgery Journal*, 3(04):e163–e166, 2017.

[41] Max Allan, Sébastien Ourselin, Steve Thompson, David J Hawkes, John Kelly, and Danail Stoyanov. Toward detection and localization of instruments in minimally invasive surgery. *IEEE Transactions on Biomedical Engineering*, 60(4):1050–1058, 2013.

[42] Siddharth Kannan, Gaurav Yengera, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Future-state predicting lstm for early surgery type recognition. *IEEE Transactions on Medical Imaging*, 39(3):556–566, 2019.

[43] Carly R Garrow, Karl-Friedrich Kowalewski, Linhong Li, Martin Wagner, Mona W Schmidt, Sandy Engelhardt, Daniel A Hashimoto, Hannes G Kenngott, Sebastian Bodenstedt, Stefanie Speidel, et al. Machine learning for surgical phase recognition: a systematic review. *Annals of surgery*, 273(4):684–693, 2021.

[44] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.

[45] Weicheng Kuo, Christian Häne, Esther Yuh, Pratik Mukherjee, and Jitendra Malik. Cost-sensitive active learning for intracranial hemorrhage detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 715–723. Springer, 2018.

[46] Mohamed Alsheakhali, Mehmet Yigitsoy, Abouzar Eslami, and Nassir Navab. Surgical tool detection and tracking in retinal microsurgery. In *Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 9415, page 941511. International Society for Optics and Photonics, 2015.

[47] 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. *CoRR*, abs/1703.10701, 2017.

[48] Stefanie Speidel, Michael Delles, Carsten Gutt, and Rüdiger Dillmann. Tracking of instruments in minimally invasive surgery for surgical skill analysis. In *International Workshop on Medical Imaging and Virtual Reality*, pages 148–155. Springer, 2006.

[49] David Bouget, Max Allan, Danail Stoyanov, and Pierre Jannin. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. *Medical image analysis*, 35:633–654, 2017.

[50] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al. Surgical data science–from concepts toward clinical translation. *Medical image analysis*, 76:102306, 2022.

[51] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

[52] Niveditha Kalavakonda, Blake Hannaford, Zeeshan Qazi, and Laligam Sekhar. Autonomous neurosurgical instrument segmentation using end-to-end learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[53] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77:157–173, 2008.

[54] Stefanie Speidel, Julia Benzko, Sebastian Krappe, Gunther Sudra, Pedram Azad, Beat Peter Müller-Stich, Carsten Gutt, and Rüdiger Dillmann. Automatic classification of minimally invasive instruments based on endoscopic image sequences. In *Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling*, volume 7261, pages 106–113. SPIE, 2009.

[55] David Bouget, Rodrigo Benenson, Mohamed Omran, Laurent Riffaud, Bernt Schiele, and Pierre Jannin. Detecting surgical tools by modelling local appearance and global shape. *IEEE Transactions on Medical Imaging*, 34(12):2603–2617, dec 2015.

[56] Markus Philipp, Anna Alperovich, Marielena Gutt-Will, Andrea Mathis, Stefan Saur, Andreas Raabe, and Franziska Mathis-Ullrich. Localizing neurosurgical instruments across domains and in the wild. In Mattias Heinrich, Qi Dou, Marleen de Bruijne, Jan Lellmann, Alexander Schläfer, and Floris Ernst, editors, *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, volume 143 of *Proceedings of Machine Learning Research*, pages 581–595. PMLR, 07–09 Jul 2021.

[57] Mark Rodrigues, Michael Mayo, and Panos Patros. Surgical tool datasets for machine learning research: a survey. *International Journal of Computer Vision*, 130(9):2222–2248, 2022.

[58] K. Jo, B. Choi, S. Choi, Y. Moon, and J. Choi. Automatic detection of hemorrhage and surgical instrument in laparoscopic surgery image. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1260–1263, Aug 2016.

[59] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, Sept 2010.

[60] Zhen-Liang Ni, Gui-Bin Bian, Xiao-Liang Xie, Zeng-Guang Hou, Xiao-Hu Zhou, and Yan-Jie Zhou. Rasnet: Segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5735–5738. IEEE, 2019.

[61] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 624–628. IEEE, 2018.

[62] Daniil Pakhomov, Vittal Premachandran, Max Allan, Mahdi Azizian, and Nassir Navab. Deep residual learning for instrument segmentation in robotic surgery. In *International Workshop on Machine Learning in Medical Imaging*, pages 566–573. Springer, 2019.

[63] Tobias Ross, Annika Reinke, Peter M Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc Filimon, Patrick Scholz, Thuy Nuong Tran, et al. Robust medical instrument segmentation challenge 2019. *arXiv preprint arXiv:2003.10299*, 2020.

[64] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[65] Daniil Pakhomov, Vittal Premachandran, Max Allan, Mahdi Azizian, and Nassir Navab. Deep residual learning for instrument segmentation in robotic surgery. *CoRR*, abs/1703.08580, 2017.

[66] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[67] Robotic instrument segmentation sub-challenge. https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/. Accessed: 2017-07-02.

[68] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[69] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792, 2014.

[70] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[71] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*, pages 1–4. IEEE, 2017.

[72] Mark B. Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[73] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[74] Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation. *CoRR*, abs/1801.05746, 2018.

[75] Zhi Tian, Tong He, Chunhua Shen, and Youliang Yan. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. 2019.

[76] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621, 2017.

[77] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, August 2004.

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[79] Bruno Silva, Bruno Oliveira, Pedro Morais, LR Buschle, Jorge Correia-Pinto, Estevão Lima, and Joao L Vilaça. Analysis of current deep learning networks for semantic segmentation of anatomical structures in laparoscopic surgery. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3502–3505. IEEE, 2022.

[80] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021.

[81] Yutong Liu, Haijiang Zhu, Mengting Liu, Huaiyuan Yu, Zihan Chen, and Jie Gao. Rolling-unet: Revitalizing mlp's ability to efficiently extract long-distance dependencies for medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 3819–3827, 2024.

[82] Jun Liu, Kunqi Li, Chun Huang, Hua Dong, Yusheng Song, and Rihui Li. Mixformer: a mixed cnn-transformer backbone for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 2024.

[83] Shengxiang Wang, Ge Li, Min Gao, Linlin Zhuo, Mingzhe Liu, Zhizhong Ma, Wei Zhao, and Xiangzheng Fu. Gh-unet: group-wise hybrid convolution-vit for robust medical image segmentation. *npj Digital Medicine*, 8(1):426, 2025.

[84] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

[85] Hulin Li, Jun Li, Hanbing Wei, Zheng Liu, Zhenfei Zhan, and Qiliang Ren. Slim-neck by gsconv: A better design paradigm of detector architectures for autonomous vehicles. *arXiv preprint arXiv:2206.02424*, 10:5, 2022.

[86] Guoping Xu, Wentao Liao, Xuan Zhang, Chang Li, Xinwei He, and Xinglong Wu. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern recognition*, 143:109819, 2023.

[87] Wenze Liu, Hao Lu, Hongtao Fu, and Zhiguo Cao. Learning to upsample by learning to sample. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6027–6037, 2023.

[88] Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, 2017.

[89] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[90] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[91] Christophe Giraud-Carrier. A note on the utility of incremental learning. *Ai Communications*, 13(4):215–223, 2000.

[92] Burr Settles. Active learning. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):1–114, 2012.

[93] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2):133–168, 1997.

[94] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

[95] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9368–9377, 2018.

[96] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.

[97] Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. Cereals-cost-effective region-based active learning for semantic segmentation. *arXiv preprint arXiv:1810.09726*, 2018.

[98] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019.

[99] Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.

[100] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2015.

[101] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[102] Ashish Kapoor, Eric Horvitz, and Sumit Basu. Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *IJCAI*, volume 7, pages 877–882, 2007.

[103] Alexander Vezhnevets, Joachim M Buhmann, and Vittorio Ferrari. Active learning for semantic segmentation with expected change. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3162–3169. IEEE, 2012.

[104] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*, pages 2372–2379. IEEE, 2009.

[105] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Discovering general-purpose active learning strategies. *arXiv preprint arXiv:1810.04114*, 2018.

[106] Suyog Dutt Jain and Kristen Grauman. Active image segmentation propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2864–2873, 2016.

[107] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9433–9443, 2020.

[108] Arantxa Casanova, Pedro O Pinheiro, Negar Rostamzadeh, and Christopher J Pal. Reinforced active learning for image segmentation. *arXiv preprint arXiv:2002.06583*, 2020.

[109] Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[110] Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 661–670, 2014.

[111] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079, 2008.

[112] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[113] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[114] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[115] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[116] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[117] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015.

[118] Josiah Aklilu and Serena Yeung. Alges: active learning with gradient embeddings for semantic segmentation of laparoscopic surgical images. In *Machine Learning for Healthcare Conference*, pages 892–911. PMLR, 2022.

[119] Haonan Peng, Shan Lin, Daniel King, Yun-Hsuan Su, Waleed M Abuzeid, Randall A Bly, Kris S Moe, and Blake Hannaford. Reducing annotating load: Active learning with synthetic images in surgical instrument segmentation. *Medical Image Analysis*, 97:103246, 2024.

[120] Luke Gessler, Lauren Levine, and Amir Zeldes. Midas loop: A prioritized human-in-the-loop annotation for large scale multilayer data. In *Proceedings of the 16th Lingusitic Annotation Workshop (LAW-XVI) within LREC2022*, pages 103–110, 2022.

[121] Jonas Jonas Boysen and Anthony Stein. Ai-supported data annotation in the context of uav-based weed detection in sugar beet fields using deep neural networks. *42. GIL-Jahrestagung, Künstliche Intelligenz in der Agrar-und Ernährungswirtschaft*, 2022.

[122] Douwe van der Wal, Iny Jhun, Israa Laklouk, Jeff Nirschl, Lara Richer, Rebecca Rojansky, Talent Theparee, Joshua Wheeler, Jörg Sander, Felix Feng, et al. Biological data annotation via a human-augmenting ai-based labeling system. *NPJ Digital Medicine*, 4(1):145, 2021.

[123] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[124] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[125] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *arXiv preprint arXiv:2212.06823*, 2022.

[126] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise, 2018.

[127] Leonardo C da Cruz, César A Sierra-Franco, Greis Francy M Silva-Calpa, and Alberto Barbosa Raposo. Enabling autonomous medical image data annotation: A human-in-the-loop reinforcement learning approach. In *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 271–279. IEEE, 2021.

[128] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T Wolf, et al. Ai-assisted human labeling: Batching for efficiency without overreliance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–27, 2021.

[129] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[130] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[131] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

[132] Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. Analysis of the isic image datasets: Usage, benchmarks and recommendations. *Medical image analysis*, 75:102305, 2022.

[133] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[134] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

[135] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[136] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[137] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016.

[138] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[139] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014.

[140] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[141] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[142] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[143] Pietro Mascagni, Deepak Alapatt, Luca Sestini, Maria S Altieri, Amin Madani, Yusuke Watanabe, Adnan Alseidi, Jay A Redan, Sergio Alfieri, Guido Costamagna, et al. Computer vision in surgery: from potential to clinical value. *npj Digital Medicine*, 5(1):163, 2022.

[144] 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. *arXiv preprint arXiv:1805.08569*, 2018.

[145] 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 *International Workshop on Computer-Assisted and Robotic Endoscopy*, pages 85–93. Springer, 2018.

[146] Luca Sestini, Benoit Rosa, Elena De Momi, Giancarlo Ferrigno, and Nicolas Padoy. A kinematic bottleneck approach for pose regression of flexible surgical instruments directly from images. *IEEE Robotics and Automation Letters*, 6(2):2938–2945, 2021.

[147] Ajay Kumar Tanwani, Pierre Sermanet, Andy Yan, Raghav Anand, Mariano Phielipp, and Ken Goldberg. The authors of use re-colorization as an auziliary task to pre-train a network for surgical instrument segmentation on the miccai 2017 dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2174–2181. IEEE, 2020.

[148] 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. *arXiv preprint arXiv:1705.08260*, 2017.

[149] 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. *International journal of computer assisted radiology and surgery*, 13:925–933, 2018.

[150] Xingtong Liu, Ayushi Sinha, Masaru Ishii, Gregory D Hager, Austin Reiter, Russell H Taylor, and Mathias Unberath. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *IEEE transactions on medical imaging*, 39(5):1438–1447, 2019.

[151] Haotian Yang and Lueder A Kahrs. Real-time coarse-to-fine depth estimation on stereo endoscopic images with self-supervised learning. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 733–737. IEEE, 2021.

[152] Samyakh Tukra and Stamatia Giannarou. Stereo depth estimation via self-supervised contrastive representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 604–614. Springer, 2022.

[153] 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. *Medical image analysis*, 77:102338, 2022.

[154] Mert Asim Karaoglu, Viktoria Markova, Nassir Navab, Benjamin Busam, and Alexander Ladikos. Ride: Self-supervised learning of rotation-equivariant keypoint detection and invariant description for endoscopy. *arXiv preprint arXiv:2309.09563*, 2023.

[155] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, Feb 2008.

[156] Tobias Rückert, Daniel Rueckert, and Christoph Palm. Methods and datasets for segmentation of minimally invasive surgical instruments in endoscopic images and videos: A review of the state of the art. *arXiv preprint arXiv:2304.13014*, 2023.

[157] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

[158] Juan Carlos Ángeles Cerón, Gilberto Ochoa-Ruiz, Leonardo Chang, and Sharib Ali. Real-time instance segmentation of surgical instruments using attention and multi-scale feature fusion. *Medical Image Analysis*, 81:102569, 2022.

[159] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[160] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[161] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, others, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[162] Wenxi Yue, Jing Zhang, Kun Hu, Yong Xia, Jiebo Luo, and Zhiyong Wang. Surgicalsam: Efficient class promptable surgical instrument segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. arXiv:2308.08746.

[163] Wenxi Yue, Jing Zhang, Kun Hu, Qiuxia Wu, Zongyuan Ge, Yong Xia, Jiebo Luo, and Zhiyong Wang. Surgicalpart-sam: Part-to-whole collaborative prompting for surgical instrument segmentation. *arXiv preprint arXiv:2312.14481*, 2024.

[164] Zhongdao Wang, Henghui Ding, and Xudong Jiang. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[165] Shu Yang, Haotian Liu, Xiaodan Zhang, Ying Shan, Xiaodong Qie, Thomas S. Huang, and Jian Yang. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[166] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[167] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23592–23601, June 2023.

[168] Wenxuan Wang, Tongtian Yue, Yisi Zhang, Longteng Guo, Xingjian He, Xinlong Wang, and Jing Liu. Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12998–13008, June 2024.

[169] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[170] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 724–732, 2016.

[171] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3367–3375, 2023.

[172] Claudia D'Ettorre, Andrea Mariani, Agostino Stilli, Ferdinando Rodriguez y Baena, Pietro Valdastri, Anton Deguet, Peter Kazanzides, Russell H Taylor, Gregory S Fischer, Simon P DiMaio, et al. Accelerating surgical robotics research: A review of 10 years with the da vinci research kit. *IEEE Robotics & Automation Magazine*, 28(4):56–78, 2021.

[173] Ming Han Lincoln Liow, Pak Lin Chin, Hee Nee Pang, Darren Keng-Jin Tay, and Seng-Jin Yeo. Think surgical tsolution-one®(robodoc) total knee arthroplasty. *SICOT-J*, 3, 2017.

[174] TRK Varma and P Eldridge. Use of the neuromate stereotactic robot in a frameless mode for functional neurosurgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 2(2):107–113, 2006.

[175] On-Road Automated Driving (ORAD) Committee. *Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems*. SAE International, 2014.

[176] Michael Yip and Nikhil Das. Robot autonomy for surgery. In *The Encyclopedia of MEDICAL ROBOTICS: Volume 1 Minimally Invasive Surgical Robotics*, pages 281–313. World Scientific, 2019.

[177] Fanny Ficuciello, Guglielmo Tamburrini, Alberto Arezzo, Luigi Villani, and Bruno Siciliano. Autonomy in surgical robots and its meaningful human control. *Paladyn, Journal of Behavioral Robotics*, 10(1):30–43, 2019.

[178] Andrew F Coburn and Zach Gage-Croll. Improving hospital patient safety through teamwork: The use of teamstepps in critical access hospitals. *challenge*, 5:7, 2011.

[179] John T James. A new, evidence-based estimate of patient harms associated with hospital care. *Journal of patient safety*, 9(3):122–128, 2013.

[180] Riccardo Muradore. Saras smart autonomous robotic assistant surgeon, 2023.

[181] Vignesh Manoj Varier, Dhruv Kool Rajamani, Nathaniel Goldfarb, Farid Tavakkolmoghaddam, Adnan Munawar, and Gregory S Fischer. Collaborative suturing: A reinforcement learning approach to automate hand-off task in suturing for surgical robots. In *2020 29th IEEE international conference on robot and human interactive communication (RO-MAN)*, pages 1380–1386. IEEE, 2020.

[182] Ben Kehoe, Gregory Kahn, Jeffrey Mahler, Jonathan Kim, Alex Lee, Anna Lee, Keisuke Nakagawa, Sachin Patil, W Douglas Boyd, Pieter Abbeel, et al. Autonomous multilateral debridement with the raven surgical robot. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1432–1439. IEEE, 2014.

[183] Chiyu Wang, João Cartucho, Daniel Elson, Ara Darzi, and Stamatia Giannarou. Towards autonomous control of surgical instruments using adaptive-fusion tracking and robot self-calibration. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2395–2401. IEEE, 2022.

[184] 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. *Frontiers in Robotics and AI*, 9:832208, 2022.

[185] Minho Hwang, Brijen Thananjeyan, Daniel Seita, Jeffrey Ichnowski, Samuel Paradis, Danyal Fer, Thomas Low, and Ken Goldberg. Superhuman surgical peg transfer using depth-sensing and deep recurrent neural networks. *arXiv preprint arXiv:2012.12844*, 2020.

[186] Florian Richter, Ryan K Orosco, and Michael C Yip. Open-sourced reinforcement learning environments for surgical robotics. *arXiv preprint arXiv:1903.02090*, 2019.

[187] Sahba Aghajani Pedram, Peter Walker Ferguson, Changyeob Shin, Ankur Mehta, Erik P Dutson, Farshid Alambeigi, and Jacob Rosen. Toward synergic learning for autonomous manipulation of deformable tissues via surgical robots: An approximate

q-learning approach. In *2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, pages 878–884. IEEE, 2020.

[188] 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)*, pages 35–42. IEEE, 2018.

[189] Daniel Seita, Sanjay Krishnan, Roy Fox, Stephen McKinley, John Canny, and Ken Goldberg. Fast and reliable autonomous surgical debridement with cable-driven robots using a two-phase calibration procedure. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6651–6658. IEEE, 2018.

[190] Shahriar Sefati, Rachel Hegeman, Iulian Iordachita, Russell H Taylor, and Mehran Armand. A dexterous robotic system for autonomous debridement of osteolytic bone lesions in confined spaces: Human cadaver studies. *IEEE Transactions on Robotics*, 38(2):1213–1229, 2021.

[191] Danying Hu, Yuanzheng Gong, Blake Hannaford, and Eric J Seibel. Semi-autonomous simulated brain tumor ablation with ravenii surgical robot using behavior tree. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3868–3875. IEEE, 2015.

[192] Elodie Constanciel, W Apoutou N'Djin, Francis Bessiere, Francoise Chavrier, Daniel Grinberg, Alexandre Vignot, Philippe Chevalier, Jean Yves Chapelon, and Cyril Lafon. Design and evaluation of a transesophageal hifu probe for ultrasound-guided cardiac ablation: simulation of a hifu mini-maze procedure and preliminary ex vivo trials. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 60(9):1868–1883, 2013.

[193] Hui Wang, Wei Kang, Thomas Carrigan, Austin Bishop, Noah Rosenthal, Mauricio Arruda, and Andrew M Rollins. In vivo intracardiac optical coherence tomography imaging through percutaneous access: toward image-guided radio-frequency ablation. *Journal of biomedical optics*, 16(11):110505–110505, 2011.

[194] Samih Taktak, Patrick Jones, Ahsanul Haq, Bhavan Prasad Rai, and Bhaskar K Somani. Aquablation: a novel and minimally invasive surgery for benign prostate enlargement. *Therapeutic Advances in Urology*, 10(6):183–188, 2018.

[195] Farshid Alambeigi, Zerui Wang, Yun-hui Liu, Russell H Taylor, and Mehran Armand. Toward semi-autonomous cryoablation of kidney tumors via model-independent deformable tissue manipulation technique. *Annals of biomedical engineering*, 46:1650–1662, 2018.

[196] Changyeob Shin, Peter Walker Ferguson, Sahba Aghajani Pedram, Ji Ma, Erik P Dutson, and Jacob Rosen. Autonomous tissue manipulation via surgical robot using learning based model predictive control. In *2019 International conference on robotics and automation (ICRA)*, pages 3875–3881. IEEE, 2019.

[197] Aleks Attanasio, Bruno Scaglioni, Matteo Leonetti, Alejandro F Frangi, William Cross, Chandra Shekhar Biyani, and Pietro Valdastri. Autonomous tissue retraction in robotic assisted minimally invasive surgery–a feasibility study. *IEEE Robotics and Automation Letters*, 5(4):6528–6535, 2020.

[198] 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. *IEEE Robotics and Automation Letters*, 5(2):2294–2301, 2020.

[199] Ameya Pore, Eleonora Tagliabue, Marco Piccinelli, Diego Dall'Alba, Alicia Casals, and Paolo Fiorini. Learning from demonstrations for autonomous soft-tissue retraction. In *2021 International Symposium on Medical Robotics (ISMR)*, pages 1–7. IEEE, 2021.

[200] Ameya Pore, Davide Corsi, Enrico Marchesini, Diego Dall'Alba, Alicia Casals, Alessandro Farinelli, and Paolo Fiorini. Safe reinforcement learning using formal verification for tissue retraction in autonomous robotic-assisted surgery. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4025–4031. IEEE, 2021.

[201] Daniele Meli, Eleonora Tagliabue, Diego Dall'Alba, and Paolo Fiorini. Autonomous tissue retraction with a biomechanically informed logic based framework. In *2021 International Symposium on Medical Robotics (ISMR)*, pages 1–7. IEEE, 2021.

[202] Ngoc Duy Nguyen, Thanh Nguyen, Saeid Nahavandi, Asim Bhatti, and Glenn Guest. Manipulating soft tissues by deep reinforcement learning for autonomous robotic surgery. In *2019 IEEE International Systems Conference (SysCon)*, pages 1–7. IEEE, 2019.

[203] Brijen Thananjeyan, Animesh Garg, Sanjay Krishnan, Carolyn Chen, Lauren Miller, and Ken Goldberg. Multilateral surgical pattern cutting in 2d orthotropic gauze with deep reinforcement learning policies for tensioning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2371–2378. IEEE, 2017.

[204] Hanh ND Le, Justin D Opfermann, Michael Kam, Sudarshan Raghunathan, Hamed Saeidi, Simon Leonard, Jin U Kang, and Axel Krieger. Semi-autonomous laparoscopic robotic electro-surgery with a novel 3d endoscope. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6637–6644. IEEE, 2018.

[205] Lijun Han, Hesheng Wang, Zhe Liu, Weidong Chen, and Xiufeng Zhang. Vision-based cutting control of deformable objects with surface tracking. *IEEE/ASME Transactions on Mechatronics*, 26(4):2016–2026, 2020.

[206] Takayuki Osa, Naohiko Sugita, and Mamoru Mitsuishi. Online trajectory planning and force control for automation of surgical tasks. *IEEE Transactions on Automation Science and Engineering*, 15(2):675–691, 2017.

[207] Simon Leonard, Kyle L Wu, Yonjae Kim, Axel Krieger, and Peter CW Kim. Smart tissue anastomosis robot (star): A vision-guided robotics system for laparoscopic suturing. *IEEE Transactions on Biomedical Engineering*, 61(4):1305–1317, 2014.

[208] Hyosig Kang and J. T. Wen. Autonomous suturing using minimally invasive surgical robots. In *Proceedings of the 2000. IEEE International Conference on Control Applications. Conference Proceedings (Cat. No.00CH37162)*, pages 742–747, Sep. 2000.

[209] Fangxun Zhong, Yaqing Wang, Zerui Wang, and Yun-Hui Liu. Dual-arm robotic needle insertion with active tissue deformation for autonomous suturing. *IEEE Robotics and Automation Letters*, 4(3):2669–2676, 2019.

[210] Sahba Aghajani Pedram, Changyeob Shin, Peter Walker Ferguson, Ji Ma, Erik P Dutson, and Jacob Rosen. Autonomous suturing framework and quantification using a cable-driven surgical robot. *IEEE Transactions on Robotics*, 37(2):404–417, 2020.

[211] Kim L Schwaner, Iñigo Iturrate, Jakob KH Andersen, Pernille T Jensen, and Thiusius R Savarimuthu. Autonomous bi-manual surgical suturing based on skills learned from demonstration. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4017–4024. IEEE, 2021.

[212] Azad Shademan, Ryan S Decker, Justin D Opfermann, Simon Leonard, Axel Krieger, and Peter CW Kim. Supervised autonomous robotic soft tissue surgery. *Science translational medicine*, 8(337):337ra64–337ra64, 2016.

[213] Sahba Aghajani Pedram, Peter Ferguson, Ji Ma, Erik Dutson, and Jacob Rosen. Autonomous suturing via surgical robot: An algorithm for optimal selection of needle diameter, shape, and path. In *2017 IEEE International conference on robotics and automation (ICRA)*, pages 2391–2398. IEEE, 2017.

[214] Claudia D'Ettorre, George Dwyer, Xiaofei Du, François Chadebecq, Francisco Vasconcelos, Elena De Momi, and Danail Stoyanov. Automated pick-up of suturing needles for robotic surgical assistance. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1370–1377. IEEE, 2018.

[215] Hamed Saeidi, Justin D Opfermann, Michael Kam, Shuwen Wei, Simon Léonard, Michael H Hsieh, Jin U Kang, and Axel Krieger. Autonomous robotic laparoscopic surgery for intestinal anastomosis. *Science robotics*, 7(62):eabj2908, 2022.

[216] Xin Ma, Chengzhi Song, Philip Waiyan Chiu, and Zheng Li. Autonomous flexible endoscope for minimally invasive surgery with enhanced safety. *IEEE Robotics and Automation Letters*, 4(3):2607–2613, 2019.

[217] J Micah Prendergast, Gregory A Formosa, Mitchell J Fulton, Christoffer R Heckman, and Mark E Rentschler. A real-time state dependent region estimator for autonomous endoscope navigation. *IEEE Transactions on Robotics*, 37(3):918–934, 2020.

[218] Huxin Gao, Weichen Fan, Liang Qiu, Xiaoxiao Yang, Zhen Li, Xiuli Zuo, Yanqing Li, Max Q-H Meng, and Hongliang Ren. Savanet: Surgical action-driven visual attention network for autonomous endoscope control. *IEEE Transactions on Automation Science and Engineering*, 2022.

[219] Tommaso Da Col, Guido Caccianiga, Michele Catellani, Andrea Mariani, Matteo Ferro, Giovanni Cordima, Elena De Momi, Giancarlo Ferrigno, and Ottavio De Cobelli. Automating endoscope motion in robotic surgery: a usability study on da vinci-assisted ex vivo neobladder reconstruction. *Frontiers in Robotics and AI*, 8:707704, 2021.

[220] Takayuki Osa, Naohiko Sugita, and Mamoru Mitsuishi. Online trajectory planning in dynamic environments for surgical task automation. In *Robotics: Science and Systems*, pages 1–9, 2014.

[221] Jur Van Den Berg, Stephen Miller, Daniel Duckworth, Humphrey Hu, Andrew Wan, Xiao-Yu Fu, Ken Goldberg, and Pieter Abbeel. Superhuman performance of surgical tasks by robots using iterative learning from human-guided demonstrations. In *2010 IEEE International Conference on Robotics and Automation*, pages 2074–2081. IEEE, 2010.

[222] Jingbin Huang, Fei Liu, Florian Richter, and Michael C Yip. Model-predictive control of blood suction for surgical hemostasis using differentiable fluid simulations. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12380–12386. IEEE, 2021.

[223] Baiquan Su, Shi Yu, Xintong Li, Yi Gong, Han Li, Zifeng Ren, Yijing Xia, He Wang, Yucheng Zhang, Wei Yao, et al. Autonomous robot for removing superficial traumatic blood. *IEEE Journal of Translational Engineering in Health and Medicine*, 9:1–9, 2021.

[224] Alexandre Krupa, Jacques Gangloff, Christophe Doignon, Michel De Mathelin, Guillaume Morel, Joël Leroy, Luc Soler, and Jacques Marescaux. Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing. *Robotics and Automation, IEEE Transactions on*, 19:842 – 853, 11 2003.

[225] Niveditha Kalavakonda. *Isosurface Visualization Using Augmented Reality for Improving Tumor Resection Outcomes*. Masters thesis, University of Washington, 2017.

[226] Allison M Okamura. Methods for haptic feedback in teleoperated robot-assisted surgery. *Industrial Robot: An International Journal*, 31(6):499–508, 2004.

[227] Hamed Saeidi, Justin D Opfermann, Michael Kam, Sudarshan Raghunathan, Simon Léonard, and Axel Krieger. A confidence-based shared control strategy for the smart tissue autonomous robot (star). In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1268–1275. IEEE, 2018.

[228] Philip H Pucher, Mikael H Sodergren, Amy C Lord, Julian Teare, Guang-Zhong Yang, and Ara Darzi. Consumer demand for surgical innovation: a systematic review of public perception of notes. *Surgical Endoscopy*, 29:774–780, 2015.

[229] Alaa Eldin Abdelaal, Prateek Mathur, and Septimiu E Salcudean. Robotics in vivo: A perspective on human–robot interaction in surgical robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:221–242, 2020.

[230] Tamas Haidegger, Stefanie Speidel, Danail Stoyanov, and Richard M Satava. Robot-assisted minimally invasive surgery—surgical robotics in the data age. *Proceedings of the IEEE*, 110(7):835–846, 2022.

[231] Iina E Aaltonen and Mikael Wahlström. Envisioning robotic surgery: Surgeons' needs and views on interacting with future technologies and interfaces. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 14(6):e1941, 2018.

[232] Mikael Wahlström, Laura Seppänen, Leena Norros, Iina Aaltonen, and Jarno Rikonen. Resilience through interpretive practice–a study of robotic surgery. *Safety Science*, 108:113–128, 2018.

[233] Uikyum Kim, Dong-Hyuk Lee, Woon Jong Yoon, Blake Hannaford, and Hyouk Ryeol Choi. Force sensor integrated surgical forceps for minimally invasive robotic surgery. *IEEE Transactions on Robotics*, 31(5):1214–1224, 2015.

[234] Fangbo Qin, Yangming Li, Yun-Hsuan Su, De Xu, and Blake Hannaford. Surgical instrument segmentation for endoscopic vision with data fusion of cnn prediction and kinematic pose. In *2019 international conference on robotics and automation (ICRA)*, pages 9821–9827. IEEE, 2019.

[235] Haonan Peng, Xingjian Yang, Yun-Hsuan Su, and Blake Hannaford. Real-time data driven precision estimator for raven-ii surgical robot end effector position. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 350–356. IEEE, 2020.

[236] Niveditha Kalavakonda, Laligam Sekhar, and Blake Hannaford. Augmented reality application for aiding tumor resection in skull-base surgery. In *2019 International Symposium on Medical Robotics (ISMR)*, pages 1–6. IEEE, 2019.

[237] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 33–40, 2006.

[238] Robin R Murphy and Debra Schreckenghost. Survey of metrics for human-robot interaction. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 197–198. IEEE, 2013.

[239] Dan R Olsen and Michael A Goodrich. Metrics for evaluating human-robot interactions. In *Proceedings of PERMIS*, volume 2003, page 4. Citeseer, 2003.

[240] Enrique Coronado, Takuya Kiyokawa, Gustavo A Garcia Ricardez, Ixchel G Ramirez-Alpizar, Gentiane Venture, and Natsuki Yamanobe. Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an industry 5.0. *Journal of Manufacturing Systems*, 63:392–410, 2022.

[241] Georgios Fagogenis, Margherita Mencattelli, Zurab Machaidze, Benoit Rosa, Karl Price, F Wu, V Weixler, Mossab Saeed, John E Mayer, and Pierre E Dupont. Autonomous robotic intracardiac catheter navigation using haptic vision. *Science robotics*, 4(29):eaaw1977, 2019.

[242] Hang Su, Juan Sandoval, Mohatashem Makhdoomi, Giancarlo Ferrigno, and Elena De Momi. Safety-enhanced human-robot interaction control of redundant robot for teleoperated minimally invasive surgery. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6611–6616. IEEE, 2018.

[243] Eleonora Tagliabue, Daniele Meli, Diego Dall'Alba, and Paolo Fiorini. Deliberation in autonomous robotic surgery: a framework for handling anatomical uncertainty. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 11080–11086. IEEE, 2022.

[244] Francesco Cursi, Valerio Modugno, Leonardo Lanari, Giuseppe Oriolo, and Petar Kormushev. Bayesian neural network modeling and hierarchical mpc for a tendon-driven surgical robot with uncertainty minimization. *IEEE Robotics and Automation Letters*, 6(2):2642–2649, 2021.

[245] Adrian Kampa. The review of reliability factors related to industrial robots. *Robotics and Automation Engineering Journal*, 3(5):1–5, 2018.

[246] Suren Kumar, Madusudanan Sathianarayanan, Sudha Garimella, Pankaj Singhal, Jason J Corso, and Venkat Krovi. Video-analytics for enhancing safety and decision-support in surgical workflows. In *Frontiers in Biomedical Devices*, volume 56000, page V001T10A020. American Society of Mechanical Engineers, 2013.

[247] N. Padoy and G. D. Hager. Human-machine collaborative surgery using learned models. In *2011 IEEE International Conference on Robotics and Automation*, pages 5285–5292, May 2011.

[248] Rishi Nayyar, Siddharth Yadav, Prabhjot Singh, and Prem Nath Dogra. Impact of assistant surgeon on outcomes in robotic surgery. *Indian journal of urology: IJU: journal of the Urological Society of India*, 32(3):204, 2016.

[249] Olivia Sgarbura and Catalin Vasilescu. The decisive role of the patient-side surgeon in robotic surgery. *Surgical endoscopy*, 24:3149–3155, 2010.

[250] Francois Lagrange, Gaelle Fiard, Clement Larose, Pascal Eschwege, and Jacques Hubert. Role and training of the bedside surgeon in robotic surgery: A survey among french urologists-in-training. *Research and Reports in Urology*, pages 17–22, 2022.

[251] Shiva Sirasala, Venkata Jammalamadaka, Moritz Spiller, Thomas Sühn, Alfredo Illanes, Axel Boese, and Michael Friebe. Collaborative robot as scrub nurse. In *Current Directions in Biomedical Engineering*, volume 7, pages 162–165. De Gruyter, 2021.

[252] E. Bauzano, B. Estebanez, I. Garcia-Morales, and V. F. Muñoz. Collaborative human–robot system for hals suture procedures. *IEEE Systems Journal*, 10(3):957–966, Sep. 2016.

[253] P. Berthet-Rayne, M. Power, H. King, and G. Yang. Hubot: A three state human-robot collaborative framework for bimanual surgical tasks based on learned models. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 715–722, May 2016.

[254] K. E. Kaplan, K. A. Nichols, and A. M. Okamura. Toward human-robot collaboration in surgery: Performance assessment of human and robotic agents in an inclusion segmentation task. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 723–729, May 2016.

[255] Luyang Zhang, Junjun Ma, Lu Zang, Feng Dong, Aiguo Lu, Bo Feng, Zirui He, Hiju Hong, and Minhua Zheng. Prevention and management of hemorrhage during a laparoscopic colorectal surgery. *Ann Laparos Endosc Surg*, 1(7):269–274, 2016.

[256] Chad G Ball and Elijah Dixon. Treatment of ongoing hemorrhage. *Cham: Springer International Publishing*, 2018.

[257] Jenna C Allers, Ahmed A Hussein, Nabeeha Ahmad, Lora Cavuoto, Joseph F Wing, Robin M Hayes, Nobuyuki Hinata, Ann M Bisantz, and Khurshid A Guru. Evaluation and impact of workflow interruptions during robot-assisted surgery. *Urology*, 92:33–37, 2016.

[258] Oscar Martinez, Christopher Murphy, Aya Bsatee, David Hugh Brown, and Ronney Abaza. Impact of surgeon-controlled suction during robotic prostatectomy to reduce dependence on bedside assistance. *Journal of Endourology*, 35(8):1163–1167, 2021.

[259] Florian Richter, Shihao Shen, Fei Liu, Jingbin Huang, Emily K Funk, Ryan K Orosco, and Michael C Yip. Autonomous robotic suction to clear the surgical field for hemostasis using image-based blood flow detection. *IEEE Robotics and Automation Letters*, 6(2):1383–1390, 2021.

[260] Giacomo De Rossi, Marco Minelli, Alessio Sozzi, Nicola Piccinelli, Federica Ferraguti, Francesco Setti, Marcello Bonfé, Cristian Secchi, and Riccardo Muradore. Cognitive robotic architecture for semi-autonomous execution of manipulation tasks in a surgical environment. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7833. IEEE, 2019.

[261] Giacomo De Rossi, Marco Minelli, Serena Roin, Fabio Falezza, Alessio Sozzi, Federica Ferraguti, Francesco Setti, Marcello Bonfè, Cristian Secchi, and Riccardo Muradore. A first evaluation of a multi-modal learning system to control surgical assistant robots via action segmentation. *IEEE Transactions on Medical Robotics and Bionics*, 3(3):714–724, 2021.

[262] Alaa Eldin Abdelaal, Jordan Liu, Nancy Hong, Gregory D Hager, and Septimiu E Salcudean. Parallelism in autonomous robotic surgery. *IEEE Robotics and Automation Letters*, 6(2):1824–1831, 2021.

[263] Jeffrey H Peters, Gerald M Fried, Lee L Swanstrom, Nathaniel J Soper, Lelan F Sillin, Bruce Schirmer, Kaaren Hoffman, Sages FLS Committee, et al. Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. *Surgery*, 135(1):21–27, 2004.

[264] Gerald M Fried, Liane S Feldman, Melina C Vassiliou, Shannon A Fraser, Donna Stanbridge, Gabriela Ghitulescu, and Christopher G Andrew. Proving the value of simulation in laparoscopic surgery. *Annals of surgery*, 240(3):518, 2004.

[265] Danying Hu. *Semi-Automation in Image-guided Robotic Brain Surgery*. PhD thesis, 2017.

[266] Danying Hu, Yuanzheng Gong, Eric J Seibel, Laligam N Sekhar, and Blake Hannaford. Semi-autonomous image-guided brain tumour resection using an integrated robotic system: A bench-top study. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 14(1):e1872, 2018.

[267] Marco Minelli, Alessio Sozzi, Giacomo De Rossi, Federica Ferraguti, Saverio Farsoni, Francesco Setti, Riccardo Muradore, Marcello Bonfè, and Cristian Secchi. Linear mpc-based motion planning for autonomous surgery. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5699–5706. IEEE, 2022.

[268] Guy Hoffman and Xuan Zhao. A primer for conducting experiments in human–robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(1):1–31, 2020.

[269] Blake Hannaford, Jacob Rosen, Diana C. W. Friedman, Hawkeye H. I. King, Phillip Roan, Lei Cheng, Daniel Glozman, Ji Ma, Sina Nia Kosari, and Lee W. White. Raven-ii: An open platform for surgical robotics research. *IEEE Transactions on Biomedical Engineering*, 60:954–959, 2013.

[270] Yun-Hsuan Su, Adnan Munawar, Anton Deguet, Andrew Lewis, Kyle Lindgren, Yangming Li, Russell H. Taylor, Gregory S. Fischer, Blake Hannaford, and Peter Kazanzides. Collaborative robotics toolkit (crtk): Open software framework for surgical robotics research. In *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*, pages 48–55, 2020.

[271] Frederick Proctor, Stephen Balakirsky, Zeid Kootbally, Thomas Kramer, Craig Schlenoff, and William Shackleford. The canonical robot command language (crcl). *Industrial Robot: An International Journal*, 2016.

[272] Peter Kazanzides, Zihan Chen, Anton Deguet, Gregory S Fischer, Russell H Taylor, and Simon P DiMaio. An open-source research kit for the da vinci® surgical system. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 6434–6439. IEEE, 2014.

[273] Zihan Chen, Anton Deguet, Russell H Taylor, and Peter Kazanzides. Software architecture of the da vinci research kit. In *2017 First IEEE International Conference on Robotic Computing (IRC)*, pages 180–187. IEEE, 2017.

[274] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[275] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992.

[276] JA Martin, Glenn Regehr, Richard Reznick, Helen Macrae, John Murnaghan, Carol Hutchison, and M Brown. Objective structured assessment of technical skill (osats) for surgical residents. *British journal of surgery*, 84(2):273–278, 1997.

[277] Rajesh Kumar, Amod Jog, Balazs Vagvolgyi, Hiep Nguyen, Gregory Hager, Chi Chiung Grace Chen, and David Yuh. Objective measures for longitudinal assessment of robotic surgery training. *The Journal of thoracic and cardiovascular surgery*, 143(3):528–534, 2012.

[278] Ignacio Oropesa, Patricia Sánchez-González, Magdalena K Chmarra, Pablo Lamata, Alvaro Fernández, Juan A Sánchez-Margallo, Frank Willem Jansen, Jenny Dankelman, Francisco M Sánchez-Margallo, and Enrique J Gómez. Eva: laparoscopic instrument tracking based on endoscopic video analysis for psychomotor skills assessment. *Surgical endoscopy*, 27:1029–1039, 2013.

[279] Sivakumar Balasubramanian, Alejandro Melendez-Calderon, and Etienne Burdet. A robust and sensitive metric for quantifying movement smoothness. *IEEE transactions on biomedical engineering*, 59(8):2126–2136, 2011.

[280] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[281] Mark R Wilson, Jamie M Poolton, Neha Malhotra, Karen Ngo, Elizabeth Bright, and Rich SW Masters. Development and validation of a surgical workload measure: the surgery task load index (surg-tlx). *World journal of surgery*, 35(9):1961–1969, 2011.

[282] Martin A Makary and Michael Daniel. Medical error—the third leading cause of death in the us. *Bmj*, 353, 2016.

[283] James G Anderson and Kathleen Abrahamson. Your health care may kill you: Medical errors. In *ITCH*, pages 13–17, 2017.

[284] Yoel Donchin, Daniel Gopher, Miriam Olin, Yehuda Badihi, M Biesky, CL Sprung, Reuven Pizov, and Shamay Cotev. A look into the nature and causes of human errors in the intensive care unit. *BMJ Quality & Safety*, 12(2):143–147, 2003.

[285] Ayako Okuyama, Cordula Wagner, and Bart Bijnen. Speaking up for patient safety by hospital-based health care professionals: a literature review. *BMC health services research*, 14(1):1–8, 2014.

[286] F Corcione, C Esposito, D Cuccurullo, A Settembre, N Miranda, F Amato, F Pirozzi, and P Caiazzo. Advantages and limits of robot-assisted laparoscopic surgery: preliminary experience. *Surgical Endoscopy and Other Interventional Techniques*, 19(1):117–119, 2005.

[287] Christos Bergeles and Guang-Zhong Yang. From passive tool holders to microsurgeons: safer, smaller, smarter surgical robots. *IEEE Transactions on Biomedical Engineering*, 61(5):1565–1576, 2013.

[288] Bernd Carsten Stahl and Mark Coeckelbergh. Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems*, 86:152–161, 2016.

[289] David Fischinger, Peter Einramhof, Konstantinos Papoutsakis, Walter Wohlkinger, Peter Mayer, Paul Panek, Stefan Hofmann, Tobias Koertner, Astrid Weiss, Antonis Argyros, et al. Hobbit, a care robot supporting independent living at home: First prototype and lessons learned. *Robotics and Autonomous Systems*, 75:60–78, 2016.

[290] Carlos Pérez-Vidal, E Carpintero, N Garcia-Aracil, José María Sabater-Navarro, José Maria Azorín, A Candela, and E Fernandez. Steps in the development of a robotic scrub nurse. *Robotics and Autonomous Systems*, 60(6):901–911, 2012.

[291] Hannah RM Pelikan, Amy Cheatle, Malte F Jung, and Steven J Jackson. Operating at a distance-how a teleoperated surgical robot reconfigures teamwork in the operating room. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–28, 2018.

[292] Eduard Fosch-Villaronga, Pranav Khanna, Hadassah Drukarch, and Bart Custers. The role of humans in surgery automation: Exploring the influence of automation on human–robot interaction and responsibility in surgery innovation. *International Journal of Social Robotics*, 15(3):563–580, 2023.

[293] Wiebe E Bijker. How is technology made?—that is the question! *Cambridge journal of economics*, 34(1):63–76, 2010.

[294] Garth H Ballantyne. The future of telerobotic surgery. In *Robotic Urologic Surgery*, pages 199–207. Springer, 2007.

[295] Tim Lane. A short history of robotic surgery. *The Annals of The Royal College of Surgeons of England*, 100(6_sup):5–7, 2018.

[296] David B Camarillo, Thomas M Krummel, and J Kenneth Salisbury Jr. Robotic technology in surgery: past, present, and future. *The American Journal of Surgery*, 188(4):2–15, 2004.

[297] Matt Novak. Telemedicine predicted in 1925, 2012.

[298] International Centre for Mechanical Sciences; International Federation for the Theory of Machines, Mechanisms, and Arthur D Alexander. Impacts of telemation on modern society. *On Theory and Practice of Robots and Manipulators*, pages 121–136, 1972.

[299] James C Rosser, Steven M Young, and Jonathan Klonsky. Telementoring: an application whose time has come. *Surgical endoscopy*, 21:1458–1463, 2007.

[300] Richard M Satava. Robotic surgery: from past to future—a personal journey. *Surgical Clinics*, 83(6):1491–1500, 2003.

[301] Richard M Satava. Surgical robotics: the early chronicles: a personal historical perspective. *Surgical Laparoscopy Endoscopy & Percutaneous Techniques*, 12(1):6–16, 2002.

[302] Abel-Jan Tasman and Heinz Stammberger. Video-endoscope versus endoscope for paranasal sinus surgery: influence on stereoacuity. *American journal of rhinology*, 12(6):389–392, 1998.

[303] IC Jourdan, E Dutson, A Garcia, T Vleugels, J Leroy, D Mutter, and J Marescaux. Stereoscopic vision provides a significant advantage for precision robotic laparoscopy. *Journal of British Surgery*, 91(7):879–885, 2004.

[304] Philip S Green, John W Hill, Joel F Jensen, and A Shah. Telepresence surgery. *IEEE Engineering in Medicine and Biology Magazine*, 14(3):324–329, 1995.

[305] Joel F Jensen and John W Hill. Advanced telepresence surgery system development. In *Medicine Meets Virtual Reality*, pages 107–117. IOS Press, 1996.

[306] Richard M Satava. Emerging technologies for surgery in the 21st century. *Archives of Surgery*, 134(11):1197–1202, 1999.

[307] Garth H Ballantyne and Fred Moll. The da vinci telerobotic surgical system: the virtual operative field and telepresence surgery. *Surgical Clinics*, 83(6):1293–1304, 2003.

[308] Jacques Marescaux, Joel Leroy, Francesco Rubino, Michelle Smith, Michel Vix, Michele Simone, and Didier Mutter. Transcontinental robot-assisted remote telesurgery: feasibility and potential applications. *Annals of surgery*, 235(4):487–492, 2002.

[309] Mehran Anvari. Robot-assisted remote telepresence surgery. In *Seminars in laparoscopic surgery*, volume 11, pages 123–128. Sage Publications Sage CA: Thousand Oaks, CA, 2004.

[310] Michael D Fabrlzio, Benjamin R Lee, David Y Chan, Daniel Stoianovici, Thomas W Jarrett, Calvin Yang, and Louis R Kavoussi. Effect of time delay on surgical performance during telesurgical manipulation. *Journal of endourology*, 14(2):133–138, 2000.

[311] Robert Thirsk, David Williams, and Mehran Anvari. Neemo 7 undersea mission. *Acta Astronautica*, 60(4-7):512–517, 2007.

[312] Ken Kamler. How i survived a zero-g robot operating room: Extreme surgeon, 2007.

[313] New Scientist. Doctors remove tumour in first zero-g surgery, 2006.

[314] Evalyn I George, COL Timothy C Brand, Jacques Marescaux, et al. Origins of robotic surgery: from skepticism to standard of care. *JSLS: Journal of the Society of Laparoendoscopic Surgeons*, 22(4), 2018.

[315] Richard M Satava. Virtual reality, telesurgery, and the new world order of medicine. *Journal of Image Guided Surgery*, 1(1):12–16, 1995.

[316] Pablo Garcia, Jacob Rosen, Chetan Kapoor, Mark Noakes, Greg Elbert, Michael Treat, Tim Ganous, Matt Hanson, Joe Manak, Chris Hasser, et al. Trauma pod: a semi-automated telerobotic surgical system. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 5(2):136–146, 2009.

[317] Diana CW Friedman, Jesse Dosher, Tim Kowalewski, Jacob Rosen, and Blake Hannaford. Automated tool handling for the trauma pod surgical robot. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 1936–1941. IEEE, 2007.

[318] Joseph R Sterbis, Eric J Hanly, Barry C Herman, Michael R Marohn, Timothy J Broderick, Samuel P Shih, Brett Harnett, Charles Doarn, and Noah S Schenkman. Transcontinental telesurgical nephrectomy using the da vinci robot in a porcine model. *Urology*, 71(5):971–973, 2008.

[319] M Cubano, BK Poulose, MA Talamini, R Stewart, LE Antosek, R Lentz, R Nibe, MF Kutka, and M Mendoza-Sagaon. Long distance telementoring: a novel tool for laparoscopy aboard the uss abraham lincoln. *Surgical endoscopy*, 13:673–678, 1999.

[320] Jacob Rosen and Blake Hannaford. Doc at a distance. *IEEE spectrum*, 43(10):34–39, 2006.

[321] Christopher Nguan, Brian Miller, Rajni Patel, Patrick PW Luke, and Christopher M Schlachta. Pre-clinical remote telesurgery trial of a da vinci telesurgery prototype. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 4(4):304–309, 2008.

[322] US Food and Drug Administration. Update: Caution with robotically-assisted surgical devices in mastectomy: Fda safety communication, 2021.

[323] Seung Yoon Yang, Kun Ho Roh, You-Na Kim, Minah Cho, Seung Hyun Lim, Taeil Son, Woo Jin Hyung, and Hyoung-Il Kim. Surgical outcomes after open, laparoscopic, and robotic gastrectomy for gastric cancer. *Annals of surgical oncology*, 24(7):1770–1777, 2017.

[324] Rebecca Randell, Natasha Alvarado, Stephanie Honey, Joanne Greenhalgh, Peter Gardner, Arron Gill, David Jayne, Alwyn Kotze, Alan Pearman, and Dawn Dowding. Impact of robotic surgery on decision making: perspectives of surgical teams. In *AMIA Annual Symposium Proceedings*, volume 2015, page 1057, 2015.

[325] Carsten N Gutt, T Oniu, Arianeb Mehrabi, A Kashfi, P Schemmer, and Markus W Büchler. Robot-assisted abdominal surgery. *Journal of British Surgery*, 91(11):1390–1397, 2004.

[326] Resad P Pasic, John A Rizzo, Hai Fang, Susan Ross, Matt Moore, and Candace Gunnarsson. Comparing robot-assisted with conventional laparoscopic hysterectomy: impact on cost and clinical outcomes. *Journal of minimally invasive gynecology*, 17(6):730–738, 2010.

[327] Annibale D'Annibale, Emilio Morpurgo, Valentino Fiscon, Paolo Trevisan, Gianna Sovernigo, Camillo Orsini, and Daniela Guidolin. Robotic and laparoscopic surgery for treatment of colorectal diseases. *Diseases of the colon & rectum*, 47(12):2162–2168, 2004.

[328] AL Rawlings, JH Woodland, RK Vegunta, and DL Crawford. Robotic versus laparoscopic colectomy. *Surgical endoscopy*, 21(10):1701–1708, 2007.

[329] Stefan Hirschauer. The manufacture of bodies in surgery. *Social studies of science*, 21(2):279–319, 1991.

[330] Robert Wilson. Teamwork in the operating room. *Human Organization*, 12(4):9–14, 1953.

[331] Joan S Ash, Marc Berg, and Enrico Coiera. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *Journal of the American Medical Informatics Association*, 11(2):104–112, 2004.

[332] David M Studdert, Troyen A Brennan, Eric J Thomas, and MM Rosenthal. What have we learned since the harvard medical practice study. *Medical Error: what do we know*, pages 3–34, 2002.

[333] Louise Hull, Sonal Arora, Rajesh Aggarwal, Ara Darzi, Charles Vincent, and Nick Sevdalis. The impact of nontechnical skills on technical performance in surgery: a systematic review. *Journal of the American College of Surgeons*, 214(2):214–230, 2012.

[334] Selwyn O Rogers Jr, Atul A Gawande, Mary Kwaan, Ann Louise Puopolo, Catherine Yoon, Troyen A Brennan, and David M Studdert. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. *Surgery*, 140(1):25–33, 2006.

[335] Atul A Gawande, Michael J Zinner, David M Studdert, and Troyen A Brennan. Analysis of errors reported by surgeons at three teaching hospitals. *Surgery*, 133(6):614–621, 2003.

[336] Homa Alemzadeh, Jaishankar Raman, Nancy Leveson, Zbigniew Kalbarczyk, and Ravishankar K Iyer. Adverse events in robotic surgery: a retrospective study of 14 years of fda data. *PloS one*, 11(4):e0151470, 2016.

[337] Marc R de Leval, Jane Carthey, David J Wright, Vernon T Farewell, James T Reason, et al. Human factors and cardiac surgery: a multicenter study. *The Journal of thoracic and cardiovascular surgery*, 119(4):661–672, 2000.

[338] Fuji Lai and Eileen Entin. Robotic surgery and the operating room team. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 49, pages 1070–1073. Sage Publications Sage CA: Los Angeles, CA, 2005.

[339] Stacey Cunningham, Amine Chellali, Isabelle Jaffre, J Classe, and Caroline GL Cao. Effects of experience and workplace culture in human-robot team interaction in robotic surgery: a case study. *International Journal of Social Robotics*, 5(1):75–88, 2013.

[340] Nyssen Anne-Sophie and Blavier Adélaïde. Verbal communication as a sign of adaptation in socio-technical systems: the case of robotic surgery. In *9th Bi-annual international conference on naturalistic decision making (NDM9)*. BCS Learning & Development, 2009.

[341] Jessica L Webster and Caroline GL Cao. Lowering communication barriers in operating room technology. *Human factors*, 48(4):747–758, 2006.

[342] Caroline GL Cao and Holly Taylor. Effects of new technology on the operating room team. *Work with computing systems*, pages 309–312, 2004.

[343] Franziska Huettner, Danuta Dynda, Michael Ryan, Jamie Doubet, and David L Crawford. Robotic-assisted minimally invasive surgery; a useful tool in resident training—the peoria experience, 2002–2009. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 6(4):386–393, 2010.

[344] Brigid M Gillespie, Wendy Chaboyer, Paula Longbottom, and Marianne Wallis. The impact of organisational and individual factors on team communication in surgery: a qualitative study. *International journal of nursing studies*, 47(6):732–741, 2010.

[345] Ty T Higuchi and Matthew T Gettman. Robotic instrumentation, personnel and operating room setup. *Atlas of robotic urologic surgery*, pages 15–30, 2011.

[346] Vincent Rousseau, Caroline Aubé, and André Savoie. Teamwork behaviors: A review and an integration of frameworks. *Small group research*, 37(5):540–570, 2006.

[347] Matthew Beane. Shadow learning: Building robotic surgical skill when approved means fail. *Administrative Science Quarterly*, 64(1):87–123, 2019.

[348] Sharon-Marie Weldon, Terhi Korkiakangas, Jeff Bezemer, and Roger Kneebone. Communication in the operating theatre. *Journal of British Surgery*, 100(13):1677–1688, 2013.

[349] Jeff Bezemer, Alexandra Cope, Gunther Kress, and Roger Kneebone. Do you have another johan? negotiating meaning in the operating theatre. *Applied Linguistics Review*, 2(2011), 2011.

[350] Andrew N Healey, Kamal Nagpal, Krishna Moorthy, and Charles A Vincent. Engineering the system of communication for safer surgery. *Cognition, Technology & Work*, 13(1):1–10, 2011.

[351] David M Gaba, Steven K Howard, and Stephen D Small. Situation awareness in anesthesiology. *Human factors*, 37(1):20–31, 1995.

[352] K Catchpole, A Mishra, A Handa, and P McCulloch. Teamwork and error in the operating room: analysis of skills and roles. *Annals of surgery*, 247(4):699–706, 2008.

[353] A Mishra, K Catchpole, T Dale, and P McCulloch. The influence of non-technical performance on technical outcome in laparoscopic cholecystectomy. *Surgical endoscopy*, 22(1):68–73, 2008.

[354] Andrew N Healey and Jonathan Benn. Teamwork enables remote surgical control and a new model for a surgical system emerges. *Cognition, Technology & Work*, 11(4):255–265, 2009.

[355] Ros Jacklin, Nick Sevdalis, Ara Darzi, and Charles Vincent. Mapping surgical practice decision making: an interview study to evaluate decisions in surgical care. *The American Journal of Surgery*, 195(5):689–696, 2008.

[356] Caroline E Zsambok and Gary Klein. *Naturalistic decision making*. Psychology Press, 2014.

[357] Rose Johnson, Kenton O'Hara, Abigail Sellen, Claire Cousins, and Antonio Criminisi. Exploring the potential for touchless interaction in image-guided interventional radiology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3323–3332, 2011.

[358] Jyoti Laxmi Mishra, David K Allen, and Alan D Pearman. Information sharing during multi-agency major incidents. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10, 2011.

[359] Helena M Mentis, Kenton O'Hara, Abigail Sellen, and Rikin Trivedi. Interaction proxemics and image use in neurosurgery. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 927–936, 2012.

[360] Renáta Elek, Tamás D Nagy, Dénes Á Nagy, Tivadar Garamvölgyi, Bence Takács, Péter Galambos, József K Tar, Imre J Rudas, and Tamás Haidegger. Towards surgical subtask automation—blunt dissection. In *2017 IEEE 21st International Conference on Intelligent Engineering Systems (INES)*, pages 000253–000258. IEEE, 2017.

[361] Michael Kranzfelder, Christoph Staub, Adam Fiolka, Armin Schneider, Sonja Gillen, Dirk Wilhelm, Helmut Friess, Alois Knoll, and Hubertus Feussner. Toward increased autonomy in the surgical or: needs, requests, and expectations. *Surgical endoscopy*, 27(5):1681–1688, 2013.

[362] Jiwon Ryu, Youngjin Moon, Jaesoon Choi, and Hee Chan Kim. A kalman-filter-based common algorithm approach for object detection in surgery scene to assist surgeon's situation awareness in robot-assisted laparoscopic surgery. *Journal of Healthcare Engineering*, 2018(1):8079713, 2018.

[363] S Swaroop Vedula, Masaru Ishii, and Gregory D Hager. Objective assessment of surgical technical skill and competency in the operating room. *Annual review of biomedical engineering*, 19:301–325, 2017.

[364] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 2429–2437, 2019.

[365] Margunn Aanestad. The camera as an actor design-in-use of telemedicine infrastructure in surgery. *Computer Supported Cooperative Work (CSCW)*, 12(1):1–20, 2003.

[366] Gary B Deutsch, Sandeep Anantha Sathyanarayana, Vikraman Gunabushanam, Nitin Mishra, Eugene Rubach, Harry Zemon, Jonathan DS Klein, and George DeNoto III. Robotic vs. laparoscopic colorectal surgery: an institutional experience. *Surgical endoscopy*, 26(4):956–963, 2012.

[367] Matthias Weigl, Sophia Antoniadis, Costanza Chiapponi, Christiane Bruns, and Nick Sevdalis. The impact of intra-operative interruptions on surgeons' perceived workload: an observational study in elective general and orthopedic surgery. *Surgical endoscopy*, 29(1):145–153, 2015.

[368] AN Healey, N Sevdalis, and CA Vincent. Measuring intra-operative interference from distraction and interruption observedin the operating theatre. *Ergonomics*, 49(5-6):589–604, 2006.

[369] Brian Davies. Robotic surgery–a personal view of the past, present and future. *International Journal of Advanced Robotic Systems*, 12(5):54, 2015.

[370] Amelia Compagni, Valentina Mele, and Davide Ravasi. How early implementations influence later adoptions of innovation: Social positioning and skill reproduction in the diffusion of robotic surgery. *Academy of Management Journal*, 58(1):242–278, 2015.

[371] Jennifer L Burke, Robin R Murphy, Michael D Coovert, and Dawn L Riddle. Moonlight in miami: Field study of human-robot interaction in the context of an urban search and rescue disaster response training exercise. *Human–Computer Interaction*, 19(1-2):85–116, 2004.

[372] Jennifer Casper and Robin R. Murphy. Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(3):367–385, 2003.

[373] Jill L Drury, Jean Scholtz, and Holly A Yanco. Awareness in human-robot interactions. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, volume 1, pages 912–918. IEEE, 2003.

[374] Catherine Durnell Cramton. The mutual knowledge problem and its consequences for dispersed collaboration. *Organization science*, 12(3):346–371, 2001.

[375] Robert E Kraut, Susan R Fussell, Susan E Brennan, and Jane Siegel. Understanding effects of proximity on collaboration: Implications for technologies to. *Distributed work (2002)*, 137, 2002.

[376] Brett Stoll, Samantha Reig, Lucy He, Ian Kaplan, Malte F Jung, and Susan R Fussell. Wait, can you move the robot? examining telepresence robot use in collaborative teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 14–22, 2018.

[377] Anton Simorov, R Stephen Otte, Courtni M Kopietz, and Dmitry Oleynikov. Review of surgical robotics user interface: what is the best way to control robotic surgery? *Surgical endoscopy*, 26(8):2117–2125, 2012.

[378] Rodney Brooks. *Flesh and machines: How robots will change us.* Vintage, 2003.

[379] Donald MacKenzie and Judy Wajcman. *The social shaping of technology.* Open University, 1999.

[380] Ellen Van Oost and Darren Reed. Towards a sociological understanding of robots as companions. In *International conference on human-robot personal relationship*, pages 11–18. Springer, 2010.

[381] Igor Pernek and Alois Ferscha. A survey of context recognition in surgery. *Medical & biological engineering & computing*, 55(10):1719–1734, 2017.

[382] Tom Vercauteren, Mathias Unberath, Nicolas Padoy, and Nassir Navab. Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions. *Proceedings of the IEEE*, 108(1):198–214, 2019.

[383] Stefan Franke, Max Rockstroh, Mathias Hofer, and Thomas Neumuth. The intelligent or: design and validation of a context-aware surgical working environment. *International Journal of Computer Assisted Radiology and Surgery*, 13:1301–1308, 2018.

[384] 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. *IEEE transactions on medical imaging*, 38(4):1069–1078, 2018.

[385] Siddika Parlak, Ivan Marsic, and Randall S. Burd. Activity recognition for emergency care using rfid. In *Proceedings of the 6th International Conference on Body Area Networks*, BodyNets '11, pages 40–46, ICST, Brussels, Belgium, Belgium, 2011. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[386] Qinxi Yu, Masoud Moghani, Karthik Dharmarajan, Vincent Schorp, William Chung-Ho Panitch, Jingzhou Liu, Kush Hari, Huang Huang, Mayank Mittal, Ken Goldberg, et al. Orbit-surgical: An open-simulation framework for learning surgical augmented dexterity. *arXiv preprint arXiv:2404.16027*, 2024.

[387] Mohamed Attia, Mohammed Hossny, Saeid Nahavandi, and Hamed Asadi. Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder. In *SMC*, pages 3373–3378. IEEE, 2017.

[388] Sebastian Bodenstedt, Max Allan, Anthony Agustinos, Xiaofei Du, Luis Garcia-Peraza-Herrera, Hannes Kenngott, Thomas Kurmann, Beat Müller-Stich, Sebastien Ourselin, Daniil Pakhomov, et al. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. *arXiv preprint arXiv:1805.02475*, 2018.

[389] Hassan Al Hajj, Mathieu Lamard, Pierre-Henri Conze, Soumali Roychowdhury, Xiaowei Hu, Gabija Maršalkaitė, Odysseas Zisimopoulos, Muneer Ahmad Dedmari, Fenqiang Zhao, Jonas Prellberg, et al. Cataracts: Challenge on automatic tool annotation for cataract surgery. *Medical image analysis*, 52:24–41, 2019.

[390] W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.

[391] Aneeq Zia, Kiran Bhattacharyya, Xi Liu, Max Berniker, Ziheng Wang, Rogerio Nespolo, Satoshi Kondo, Satoshi Kasai, Kousuke Hirasawa, Bo Liu, et al. Surgical tool classification and localization: results and methods from the miccai 2022 surgtoolloc challenge. *arXiv preprint arXiv:2305.07152*, 2023.

[392] Surgical robot instrument pose estimation. https://www.synapse.org/#!Synapse:syn51471789/wiki/622255. Accessed: 2023-07-11.

[393] Charles Bibby and Ian Reid. Robust real-time visual tracking using pixel-wise posteriors. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 831–844, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[394] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[395] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*, pages 850–865. Springer, 2016.

[396] Loubna Bouarfa, Oytun Akman, Armin Schneider, Pieter P. Jonker, and Jenny Dankelman. In-vivo real-time tracking of surgical instruments in endoscopic video. *Minimally Invasive Therapy & Allied Technologies*, 21(3):129–134, 2012.

[397] Austin Reiter, Peter K Allen, and Tao Zhao. Appearance learning for 3d tracking of robotic surgical tools. *The International Journal of Robotics Research*, 33(2):342–356, 2014.

[398] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[399] Ange Lou, Kareem Tawfik, Xing Yao, Ziteng Liu, and Jack Noble. Min-max similarity: A contrastive semi-supervised deep learning network for surgical tools segmentation. *IEEE Transactions on Medical Imaging*, 2023.

[400] Qimin Cheng, Qian Zhang, Peng Fu, Conghuan Tu, and Sen Li. A survey and analysis on automatic image annotation. *Pattern Recognition*, 79:242–259, 2018.

[401] Myasar Mundher Adnan, Mohd Shafry Mohd Rahim, Amjad Rehman, Zahid Mehmood, Tanzila Saba, and Rizwan Ali Naqvi. Automatic image annotation based on deep learning models: a systematic review and future challenges. *IEEE Access*, 9:50253–50264, 2021.

[402] Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. Cereals - cost-effective region-based active learning for semantic segmentation. In *British Machine Vision Conference*, 2018.

[403] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5971–5980, 2019.

[404] Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning. *arXiv: Machine Learning*, 2017.

[405] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113:113–127, 2015.

[406] Ashish Kapoor, Eric Horvitz, and Sumit Basu. Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *International Joint Conference on Artificial Intelligence*, 2007.

[407] Alexander Vezhnevets, Joachim M. Buhmann, and Vittorio Ferrari. Active learning for semantic segmentation with expected change. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3162–3169, 2012.

[408] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

[409] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607, 2021.

[410] Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. Similar: Submodular information measures based active learning in realistic scenarios. *Advances in Neural Information Processing Systems*, 34:18685–18697, 2021.

[411] Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. Prism: A rich class of parameterized submodular information measures for guided data subset selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10238–10246, 2022.

[412] Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with fisher embeddings. *Advances in Neural Information Processing Systems*, 34:8927–8939, 2021.

[413] Mohamed-Rafik Bouguelia, Slawomir Nowaczyk, KC Santosh, and Antanas Verikas. Agreeing to disagree: Active learning with noisy labels without crowdsourcing. *International Journal of Machine Learning and Cybernetics*, 9:1307–1319, 2018.

[414] Christopher H Lin, M Mausam, and Daniel S Weld. Re-active learning: Active learning with relabeling. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[415] Chicheng Zhang and Kamalika Chaudhuri. Active learning from weak and strong labelers. *Advances in Neural Information Processing Systems*, 28, 2015.

[416] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172, 2018.

[417] Qiaoning Zhang, Matthew L Lee, and Scott Carter. You complete me: Human-ai teams and complementary expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2022.

[418] Raymond R Bond, Tomas Novotny, Irena Andrsova, Lumir Koc, Martina Sisakova, Dewar Finlay, Daniel Guldenring, James McLaughlin, Aaron Peace, Victoria McGilligan, et al. Automation bias in medicine: the influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. *Journal of electrocardiology*, 51(6):S6–S11, 2018.

[419] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[420] Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022.

[421] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.

[422] Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018.

[423] Alexey Shvets, Alexander Rakhlin, Alexandr A. Kalinin, and Vladimir I. Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. *arXiv preprint arXiv:1803.01207*, 2018.

[424] Daniil Pakhomov, Vittal Premachandran, Max Allan, Mahdi Azizian, and Nassir Navab. Deep residual learning for instrument segmentation in robotic surgery. *arXiv preprint arXiv:1703.08580*, 2017.

[425] Faizal A Haji, David Rojas, Ruth Childs, Sandrine de Ribaupierre, and Adam Dubrowski. Measuring cognitive load: performance, mental effort and simulation task complexity. *Medical education*, 49(8):815–827, 2015.

[426] David W Aha. Goal reasoning: Foundations, emerging applications, and prospects. *AI Magazine*, 39(2):3–24, 2018.

[427] Michele Ginesi, Daniele Meli, Andrea Roberti, Nicola Sansonetto, and Paolo Fiorini. Autonomous task planning and situation awareness in robotic surgery. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3144–3150. IEEE, 2020.

[428] Andy S Ding, Alexander Lu, Zhaoshuo Li, Deepa Galaiya, Masaru Ishii, Jeffrey H Siewerdsen, Russell H Taylor, and Francis X Creighton. Automated extraction of anatomical measurements from temporal bone ct imaging. *Otolaryngology–Head and Neck Surgery*, 167(4):731–738, 2022.

[429] Marco Minelli, Alessio Sozzi, Giacomo De Rossi, Federica Ferraguti, Francesco Setti, Riccardo Muradore, Marcello Bonfè, and Cristian Secchi. Integrating model predictive control and dynamic waypoints generation for motion planning in surgical scenario. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3157–3163. IEEE, 2020.

Appendix A

NEUROID DATASET FRAMES

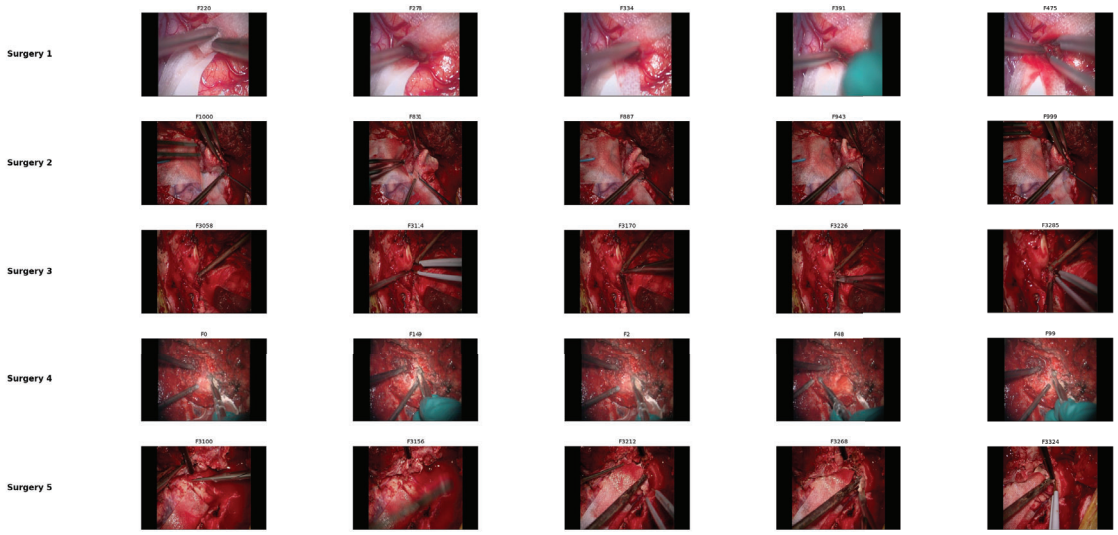


Figure A.1: Sample frames from 5 videos in the NeuroID Dataset. These indicate the diversity in surgical conditions for the procedures recorded at the Harborview Medical Center - Seattle.