

UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA



Università
di Genova

University of Modena and Reggio Emilia
Department of Sciences and Methods for Engineering

University of Genoa

PhD of National Interest in Robotics and Intelligent Machines
Curriculum: Healthcare and Wellness of Persons
XXXVIII Cycle

Toward Autonomous Soft-Tissue Minimal Invasive Surgery: Integrating Learning, Reasoning, Perceiving for Adaptive Surgical Robot Autonomy

Author:
Gabriele Furnari

Supervisor:
Federica Ferraguti

PhD Coordinator:
Antonio Sgorbissa

October 2025

Copyright© 2025 by Gabriele Furnari
Reggio Emilia, Italy, October 2025

Contents

List of Publications 6

1 Introduction **7**
1.1 Motivation and Context 7
1.2 Related Works 8
1.2.1 Autonomous Robotics in Surgery 8
1.2.2 Deep Learning and Imitation Learning for Surgical Skill 8
1.2.3 Soft Tissue Deformation and Tracking 9
1.3 Gaps in the literature 9
1.4 Goals of the Thesis 10
1.5 Challenges in Autonomous Soft-Tissue Surgery 10
1.6 Approach: Learning, Reasoning, and Perceiving 11
1.7 Thesis Contributions 12

2 Learning from Demonstrations: Imitating and Taking Decisions Through Experience **13**
2.1 Introduction: Learning from Demonstrations in surgical environment 13
2.1.1 Deep Imitation Learning in Robotics 15
2.1.2 Representation Learning for Control 16
2.1.3 Gap Analysis and Contributions 17
2.1.4 Structure of the Chapter 17
2.2 Stage I: Learning from Demonstrations in a Virtual Environment 18
2.2.1 Dataset Creation 19
2.2.2 Imitation Learning Model 24
2.2.3 Results in the Virtual Environment 28
2.2.4 Discussion and Evolution Toward Real Environments 29
2.3 Stage II: Learning from Demonstrations on Robotic Phantoms 30
2.3.1 Introduction and Motivations 30
2.3.2 Methods Overview 31
2.3.3 Dataset Creation 32
2.3.4 Model Architecture 34
2.3.5 Training of the model 35
2.4 Results and Experiments on Physical Phantoms 36
2.4.1 Quantitative Results 36
2.4.2 Qualitative Observations 37
2.4.3 Future Experimental Steps 37
2.4.4 Considerations and Future Steps 37
2.5 Conclusion: Learning from Demonstrations, Imitating and Taking Decisions Through Experience 39

3 From Learned Decisions to Surgical Reasoning: Planning for Selective Clamping 45
3.1 Introduction: Surgical Reasoning . . . 45
3.2 Problem Definition: Selective Clamping in Partial Nephrectomy . . . 46
3.3 Proposed System Architecture . . . 47
3.3.1 Clamping points computation . . . 48
3.3.2 Robotic Execution of Selective Clamping . . . 54
3.4 Experimental Validation . . . 57
3.4.1 Validation with different CTA . . . 57
3.5 Robotic execution of selective clamping . . . 57
3.6 Discussion and Future Directions . . . 58
3.7 Conclusions: Toward Contextual Autonomy . . . 59

4 Learning to Perceive: Understanding and Tracking Soft Tissue Deformation 61
4.1 Introduction: Percieving in Surgical Environment . . . 61
4.2 The Need for Understanding Soft Tissue Deformation . . . 61
4.2.1 Releted works and gaps in the literature . . . 62
4.2.2 Position within the Autonomous Surgical Framework . . . 64
4.2.3 Testing Deformation Tracking in Realistic Phantom Experiments . . . 65
4.2.4 Structure of the Chapter . . . 65
4.3 Methodological Framework . . . 66
4.3.1 Overview . . . 66
4.3.2 Non-rigid Deformation . . . 66
4.3.3 Rigid Registration . . . 69
4.4 Evaluation Metrics . . . 71
4.4.1 Silhouette Consistency . . . 71
4.4.2 Flow Consistency . . . 71
4.4.3 Depth Consistency . . . 71
4.4.4 Photometric Consistency . . . 72
4.4.5 Cutting Path Consistency . . . 72
4.5 Conclusion: Percieving modalities and future directions . . . 73
4.5.1 The Pivotal Role of Perceiving in the LRPA Paradigm . . . 74
4.5.2 Limitations and Areas for Future steps . . . 74
4.6 Future Steps in Perception Roadmap . . . 75

5 From Awareness to Automation: Integrating Learning, Reasoning, and Perception 77
5.1 Introduction: Bridging Perception, Reasoning, and Action . . . 77
5.2 Integration of Deformation Awareness into Surgical Reasoning . . . 77
5.3 Deformation-Aware Selective Clamping: Motivation and Framework . . . 78
5.3.1 Introduction and motivations . . . 78
5.3.2 The Dynamics of Arterial Deformation . . . 79
5.3.3 Implementation of Deformation-Aware Selective Clamping . . . 79
5.3.4 Conclusions and Next Steps . . . 80
5.4 Deformation-Aware Tumor Cutting and Path Following . . . 83
5.4.1 Introduction and motivations . . . 83
5.4.2 Experimental Setup and Path Definition . . . 83
5.4.3 Tool Orientation Policy Under RCM Constraints . . . 84
5.4.4 Experiments . . . 86
5.4.5 Discussion: Force Control and Justification . . . 87
5.4.6 Conclusion and Future Work . . . 87
5.5 Conclusion: From Awareness to Automation . . . 88

6 Conclusion **91**
6.1 Thesis Synthesis and Foundational Achievement 91
6.2 Summary and Contribution Mapping 91
6.2.1 Chapter 2: Learning from Demonstrations 92
6.2.2 Chapter 3: Surgical Reasoning for Selective Clamping 92
6.2.3 Chapter 4: Percieving - Soft Tissue Deformation Tracking 93
6.2.4 Chapter 5: Acting – Deformation-Aware Clamping, Tumor Cutting, and Surgical Tool Development 93
6.3 Limitations and Critical Analysis 93
6.3.1 Limitations of Imitation Learning 94
6.3.2 Limitations of Deformation Tracking 94
6.3.3 Limitations of Deformation-Aware Clamping and Cutting 94
6.3.4 Limitations in the LRPA Paradigm 94
6.4 Future Work and Research Roadmap 95
6.4.1 Future Work on Imitation Learning 95
6.4.2 Future Work on Clamping and Deformation Applications 95
6.4.3 Future Work on the LRPA Paradigm 95
6.5 Final Remarks 95

List of Publications

The following publications were produced during the course of this PhD:

- Furnari, G., Minelli, M., Puliatti, S., Micali, S., Secchi, C., Ferraguti, F. (2024, July). Selective Clamping for Robot-Assisted Surgical Procedures. In 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1-7). IEEE. **Awarded as: EMBC'24 Sudent Best Paper Competition: 3rd Place Winner**

- Furnari, G., Secchi, C., Ferraguti, F. (2025). Sequence-based imitation learning for surgical robot operations. Artificial Intelligence Surgery, 5(1), 103-115.

- Furnari, G., Hofman, J., De Backer, P., Ferraguti, F. (2025). Vision-Based Real-Time Estimation of Surgical Tissues Deformation from Video. 2025 I-RIM Conference, 121–123.

- Psychogyios, D., Colleoni, E., Van Amsterdam, B., Li, C. Y., Huang, S. Y., Li, Y., ... Stoyanov, D. (2023). Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. arXiv preprint arXiv:2401.00496.

- Furnari, G., Bigi, F., Secchi, C., Ferraguti, F. Domain-Adaptive Visuomotor Transformer for Learning Surgical Robot Trajectories. To be submitted to Transactions on Robotics: Special issue on Robot-Assisted Medical Imaging.

- Furnari, G., Iovene, E., Ferraguti, F. Vision-Driven Soft-Tissue Deformation Tracking for Adaptive Surgical Robot Behavior. To be submitted to Transactions on Robotics: Special issue on Robot-Assisted Medical Imaging.

Chapter 1

Introduction

1.1 Motivation and Context

Over the last three decades, robot-assisted surgery (RAS) has transformed several surgical specialties by enabling minimally invasive procedures with enhanced dexterity, tremor filtering and improved visualization. Systems such as the da Vinci platform have established a paradigm in which the robot acts as a teleoperated tool, fully dependent on the surgeon's continuous control input. Despite their wide clinical adoption, these systems operate at low levels of autonomy, typically providing motion scaling and ergonomic advantages rather than independent decision-making or task execution. Recent roadmaps and reviews on surgical robotics have argued that the next major leap will consist in increasing levels of autonomy, from pure teleoperation towards semi-autonomous and fully autonomous robotic systems [1].

The move towards autonomous surgical robotics (ASR) is driven by several converging needs. First, complex procedures demand sustained concentration and fine motor control over long periods, making surgeon fatigue a non-negligible risk factor. Second, there is substantial variability in surgical outcomes even among experienced surgeons, with performance depending on training, experience, and case complexity. Third, access to high-quality surgery is unevenly distributed worldwide, with many regions lacking specialized expertise. Autonomous or semi-autonomous systems that can replicate expert-level performance could improve consistency, reduce complication rates, and support wider access to advanced procedures. Early demonstrations, such as supervised autonomous suturing and anastomosis on soft tissue, have shown that robots can perform well-defined tasks with performance comparable to or exceeding that of expert surgeons in controlled scenarios [2].

However, most clinically deployed systems remain limited to rigid or quasi-rigid scenarios in which anatomy is stable or only slowly changing. The challenge becomes far more pronounced in soft-tissue surgery, where organs continuously deform due to respiration, physiological motion, gravity and direct tool interaction. Robotic-assisted partial nephrectomy (RAPN) in urology is an emblematic example: the surgeon must identify tumor margins, plan selective vascular clamping, and perform precise resections on a highly deformable kidney. Achieving autonomy in such tasks requires more than precise kinematics; it demands real-time understanding of tissue deformation, context-aware decision-making, and the ability to adapt trajectories as anatomy evolves.

This thesis is motivated by the gap between learned surgical skill and adaptive, perceptive execution in deformable environments. While imitation learning methods can capture expert demonstrations, they typically assume that the environment encountered at execution time is similar to the training conditions. In soft-tissue surgery, this assumption is untenable. The central focus of this work is therefore to develop methods that integrate learning from demonstrations, geometric reasoning and deformation-aware perception in order to achieve adaptive automation of soft-tissue surgical subtasks. In particular we used as case study two task: kidney

tumor removal and selective clamping in RAPN.

1.2 Related Works

1.2.1 Autonomous Robotics in Surgery

Autonomy in surgical robotics can be structured along levels that range from no autonomy (manual tools) to full autonomy, in which the robot can independently perceive, plan, and execute complete procedures. Recent surveys have proposed taxonomies that map existing systems to different levels of autonomy, highlighting that the vast majority of FDA-cleared surgical robots still operate at low autonomy—often limited to robot assistance and task-level automation [3]. Even when higher levels of autonomy are demonstrated, they are usually confined to specific subtasks such as suturing or tissue retraction in controlled settings.

A landmark contribution in this direction is the Smart Tissue Autonomous Robot (STAR), which demonstrated supervised autonomous soft-tissue suturing and anastomosis using 3D and near-infrared imaging to track tissue deformation and guide needle placement [4]. Subsequent works have extended autonomy to structured tasks such as knot tying, peg transfer, and simplified resection, often in benchtop settings. More recently, hierarchical models for autonomous surgery, such as the Surgical Robot Transformer (SRT and SRT-H), have sought to learn policies across multiple procedures using large-scale video and kinematic datasets, pointing towards general visuomotor policies for surgery [5].

Despite these advances, clinically deployed systems remain predominantly teleoperated. Autonomous behaviors are either confined to highly constrained environments or evaluated ex vivo. Ethical, regulatory and safety concerns further limit direct deployment of fully autonomous systems in the operating room, making it essential to develop intermediate frameworks that integrate autonomy in a controlled, verifiable manner [6].

1.2.2 Deep Learning and Imitation Learning for Surgical Skill

Deep learning has had a substantial impact on surgical data science, leading to significant progress in tasks such as surgical phase recognition, instrument detection, and skill assessment. Convolutional and recurrent neural networks have been successfully applied to recognize surgical phases in laparoscopic procedures, using endoscopic video alone or in conjunction with tool signals. EndoNet introduced a convolutional architecture for simultaneous phase and tool presence recognition in cholecystectomy, and subsequent works have extended this line of research with spatio-temporal models, attention mechanisms, and self-supervised representations [7].

Imitation learning (IL) has emerged as a key paradigm for acquiring control policies in robotics from expert demonstrations. Surveys on imitation learning distinguish between behavioral cloning, which treats the problem as supervised learning from state–action pairs, and inverse reinforcement learning, which infers an underlying reward function that explains the expert behavior [8]. In the context of robotic manipulation, deep imitation learning has been used to learn complex multi-step behaviors from visual observations, demonstrating good performance on a variety of tasks [9].

In surgical robotics, IL has been applied to learn policies for subtasks such as needle passing, tissue retraction and suturing from expert kinematic and video data. Recent works such as Surgical Robot Transformer leverage large-scale datasets to learn transformer-based visuomotor policies that generalize across tasks and instruments [10]. sequence-based models, including those using recurrent and attention mechanisms, have been proposed to capture temporal dependencies in complex surgical trajectories [11]. Compared to reinforcement learning, IL is more suitable for surgical applications because it does not require trial-and-error exploration in the real environment, which would be ethically infeasible and unsafe.

Nevertheless, most existing IL approaches are limited in their ability to generalize across domains, such as different patients, phantoms or anatomical variations, and typically assume that the geometry encountered at test time is similar to the training conditions. This is particularly problematic in soft-tissue surgery, where patient-specific anatomy and tool-induced deformation can significantly alter the environment.

1.2.3 Soft Tissue Deformation and Tracking

Modeling and tracking soft tissue deformation has been an active area of research in image-guided surgery and computer-assisted interventions. Early approaches relied on biomechanical models, often based on the finite element method (FEM), to simulate soft tissue response to applied forces [12]. While FEM provides physically grounded estimates, it is computationally expensive and requires accurate knowledge of tissue properties and boundary conditions, which can be difficult to obtain intraoperatively.

In parallel, computer vision methods have been developed to track soft tissue deformation directly from images. Stoyanov and colleagues combined stereo feature tracking with constrained surface models to recover soft tissue deformation in robotic minimally invasive surgery [13]. Subsequent works have explored optical flow, non-rigid registration, and piecewise affine models to estimate deformation from monocular or stereo laparoscopic video, often under challenging conditions such as low texture, smoke and instrument occlusion [14].

More recent reviews on tracking and mapping in medical computer vision have emphasized the importance of real-time deformation estimation for tasks such as margin delineation, motion compensation, and autonomous navigation [15]. However, much of this work remains observational or diagnostic: deformation estimates are used for visualization, navigation or guidance rather than being tightly integrated into closed-loop robotic control or task-level reasoning. In other words, deformation is often perceived, but rarely acted upon by autonomous controllers.

In summary, the current literature provides strong foundations in autonomous surgical subtasks, learning from expert demonstrations, and soft tissue deformation tracking. Yet, these components have largely been studied in isolation. There remains a need for integrated frameworks that use deformation awareness not only to observe but to adapt and correct robotic actions in real time.

1.3 Gaps in the literature

Based on the state of the art, several critical gaps can be identified that motivate the contributions of this thesis.

First, there is a gap in domain generalization for imitation learning policies in surgery. Policies trained on a specific phantom or anatomical model often fail when confronted with new geometries, different tumor shapes, or altered scene conditions. Visual domain shift, geometric variation, and differences in tool–tissue interaction lead to degraded performance or even catastrophic failure. This gap is especially acute for tasks such as tumor cutting, where the trajectory must adapt to patient-specific anatomy while maintaining strict safety margins.

Second, there is a gap in perceptive reasoning: the integration of real-time deformation tracking into high-level surgical planning and control. Existing deformation estimation methods typically do not feed into the update of critical surgical targets, such as selective clamping points on renal arteries or cutting paths on tumor margins. For true autonomy, the robot must not only estimate deformation but also use it to update target points, trajectories and decisions continuously. This includes, for example, recomputing the position of clamping points as the kidney is mobilized, or adjusting a cutting path as tissue is retracted or deformed by tool contact.

Third, there is a gap in automating complex, therapeutically relevant tasks that require multi-tool coordination and precise interaction with deformable anatomy. While many works

focus on simplified tasks such as peg transfer, relatively few address multi-step procedures such as selective vascular clamping or non-linear tumor resection, which are central to RAPN and other soft-tissue surgeries.

Finally, there is a hardware interface gap: research platforms often lack actuated, sensorized surgical tools that can be seamlessly integrated with open-architecture robots (such as KUKA or UR arms). Without such tools, it is difficult to conduct realistic experiments that involve controlled cutting, grasping and clamping, and to collect the high-fidelity data necessary for training and validating learning-based controllers.

This thesis is designed to address these gaps by combining domain-adaptive imitation learning, real-time deformation tracking, surgical reasoning, and custom actuation into a unified framework.

1.4 Goals of the Thesis

The overarching goal of this thesis is to establish a methodology for adaptive automation of soft-tissue surgical subtasks by integrating visual perception of deformation, expert knowledge acquired through imitation learning, and geometric reasoning about anatomy and surgical objectives.

More concretely, the thesis pursues the following specific objectives:

- Develop a domain-adaptive imitation learning framework for generalizing tumor cutting trajectories across different phantom geometries and tumor configurations, while maintaining clinically meaningful accuracy.
- Implement an end-to-end, real-time soft tissue deformation tracking pipeline tailored to surgical scenes, using RGB-D or stereo imaging to estimate a deformation field $\mathbf{D}(\mathbf{x}, t)$ that captures non-rigid motion of the kidney and tumor phantoms.
- Design and fabricate a novel actuated surgical tool with a precision cam–piston mechanism suitable for robotic mounting, enabling controlled cutting, grasping and future clamping operations.
- Integrate the deformation-aware perceptual framework into selective vascular clamping for RAPN, by updating clamping points along arterial centerlines as the kidney deforms.
- Validate the integrated perception-to-action system on physical silicone organs derived from CT data, focusing on deformation-aware tumor cutting and path following under externally induced and self-induced deformation.

These objectives are addressed through a combination of algorithm development, mechanical design, and experimental validation on realistic phantoms.

1.5 Challenges in Autonomous Soft-Tissue Surgery

Achieving autonomous behavior in soft-tissue surgery poses numerous challenges that span data, environment, safety and hardware.

From a data perspective, acquiring high-quality demonstrations is difficult and costly. Expert surgeon time is limited, and recording intraoperative data raises ethical and legal concerns related to patient privacy and data sharing. Furthermore, annotating continuous, multi-instrument procedures at fine temporal and spatial resolution is extremely labor-intensive, especially when complex actions such as cutting, clamping or suturing must be labeled accurately.

Environmentally, soft tissues exhibit non-linear, history-dependent and heterogeneous deformation under load. Organs such as the kidney change shape due to respiration, patient

positioning and tool interaction, leading to continuous scene evolution. Visual conditions in minimally invasive surgery are also challenging: lighting varies, specular reflections and smoke occur, fluids obscure the view, and instruments frequently occlude regions of interest. Geometric variability between patients further complicates the design of generalizable models.

From a safety and reliability standpoint, there is no universally agreed quantitative definition of safety for autonomous surgical actions. Errors in perception or control can accumulate over time, leading to covariate shift and potentially unsafe states. Unlike industrial robotics, where safety can often be guaranteed by physical separation or guarding, surgical robots operate in direct contact with delicate tissues. Any autonomous behavior must therefore be constrained and verifiable.

Hardware limitations add further complexity. Robotic arms exhibit calibration errors, joint backlash and model uncertainties that affect the accuracy of tool positioning. Camera–robot calibration is non-trivial and susceptible to drift. Many research setups lack integrated force or tactile sensing at the end-effector, which is crucial for regulating contact forces and detecting unsafe interactions in soft-tissue tasks.

The approach proposed in this thesis is designed with these challenges in mind, seeking to advance autonomy in a stepwise, experimentally grounded manner.

1.6 Approach: Learning, Reasoning, and Perceiving

The methodology adopted in this thesis is organized around three functional pillars: learning, perceiving and reasoning, tightly coupled with a dedicated actuation interface.

Learning provides the foundation by capturing expert behavior through imitation learning. In Chapter 2, policies are trained to reproduce complex surgical subtasks such as tumor cutting using synchronized video and kinematic data recorded from expert demonstrations in both simulated and physical environments. These policies encode a blueprint for how tasks should be executed under nominal conditions.

Perceiving provides adaptive input by estimating the real-time state of the environment. Chapter 4 develops a soft tissue deformation tracking pipeline that uses RGB-D or stereo vision to estimate a deformation field $\mathbf{D}(\mathbf{x}, t)$ for the kidney and tumor phantoms. This field allows the system to update the positions of anatomical structures and pre-planned trajectories as the organ deforms, bridging the gap between static planning and dynamic execution.

Reasoning forms the decision layer that uses the deformation information to adapt high-level plans. In the context of selective clamping, deformation-aware reasoning updates clamping points along arterial centerlines as the kidney moves, maintaining alignment between the intended vascular targets and their current locations. For tumor cutting, the same principle is used to deform the cutting path so that it continues to follow the tumor–kidney boundary despite ongoing deformation.

enabling both cutting and grasping operations on the phantoms.

The thesis is structured as follows. Chapter 2 focuses on learning from demonstrations and domain adaptation for tumor cutting trajectories. Chapter 3 establishes the reasoning layer of the proposed framework, formalizing the anatomical and geometric decision-making required for selective vascular clamping and defining the surgical targets that must be preserved and updated during autonomous exe

Chapter 4 details the soft tissue deformation tracking framework and its validation on silicone phantoms. Chapter 5 integrates these components, demonstrating deformation-aware clamping and tumor cutting. Finally, Chapter 6 summarizes the contributions and outlines directions for future work.

1.7 Thesis Contributions

The main contributions of this thesis can be summarized as follows:

- A domain-adaptive imitation learning framework for complex surgical trajectories, demonstrated on tumor cutting tasks across varying phantom geometries.
- The first formulation and experimental demonstration of deformation-aware selective clamping on physical phantoms, in which clamping points are continuously updated based on real-time tissue deformation estimates.
- The design, implementation and experimental validation of a custom actuated surgical tool, providing an open-architecture end-effector for autonomous soft-tissue manipulation on KUKA/UR robots.
- A physical validation of an integrated perception-to-action pipeline for deformation-aware tumor cutting on CT-derived silicone kidney–tumor phantoms.
- The creation of datasets for dual-tool kidney tumor removal, including cutting and grasping under realistic visual and deformation conditions.
- The introduction of a comprehensive framework that unifies learning from demonstrations, geometric reasoning and deformation-aware perception, providing a structured pathway toward safe and adaptive surgical autonomy.

Chapter 2

Learning from Demonstrations: Imitating and Taking Decisions Through Experience

2.1 Introduction: Learning from Demonstrations in surgical environment

Robotic assistance has fundamentally transformed the landscape of urologic surgery over the past two decades. Platforms such as the da Vinci Surgical System have become the gold standard for complex reconstructive procedures, most notably Robot-Assisted Partial Nephrectomy (RAPN). RAPN is currently the preferred treatment modality for localized renal masses, aiming to achieve the "trifecta" of surgical success: negative margins (oncologic control), maximal preservation of healthy parenchyma (functional preservation), and minimal perioperative complications. Despite the significant ergonomic and visual benefits provided by stereoscopic teleoperation, clinical outcomes remain heavily dependent on the individual surgeon's expertise. Large multi-center studies and learning-curve analyses consistently show that variability in technical skill directly correlates with inconsistencies in warm ischemia time, positive surgical margin rates, and complication profiles.

The next frontier in surgical robotics is the transition from pure teleoperation to supervised autonomy. In this paradigm, the robot is not merely a passive tool but an active partner capable of autonomously planning and executing well-defined, high-frequency, or high-precision sub-tasks under human supervision. This shift promises to democratize expert-level care by standardizing critical maneuvers, reducing surgeon cognitive load, and enabling telesurgery in settings where communication latency makes direct teleoperation unstable.

Recent milestones have demonstrated that autonomous systems can, under controlled conditions, match or even surpass human surgeons for specific tasks. The Smart Tissue Autonomous Robot (STAR) of Shademan et al. showed that supervised autonomous soft-tissue anastomosis in porcine bowel can achieve more consistent suture spacing and higher burst pressures than expert surgeons, using plenoptic 3D vision and near-infrared fluorescence to guide an engineered path planner [16].

More recently, the Surgical Robot Transformer (SRT) and its hierarchical extension SRT-H by Kim et al. demonstrated that Transformer-based imitation learning can automate fundamental skills and, in the case of SRT-H, perform ex vivo cholecystectomy in porcine models with a hierarchical language-conditioned policy that sequences low-level skills into complete procedural steps.

These works strongly suggest that soft-tissue autonomy is technically feasible at least for carefully defined subtasks and experimental procedures.

However, extending autonomy to tumor resection in RAPN presents a challenge of a different magnitude. Unlike suturing or knot-tying—largely geometric tasks involving well-defined tool–needle interactions—tumor resection is a dynamic, irreversible subtractive process. It requires continuous estimation of the tumor boundary within highly deformable tissue, precise instrument–tissue interaction to avoid vascular injury, and real-time adaptation to patient-specific viscoelastic properties. Preoperative 3D reconstructions and augmented-reality overlays are increasingly available, but intraoperative tissue deformation quickly invalidates rigid registrations, especially once hilar clamping, mobilization, and tumor traction are applied. This context demands real-time, closed-loop control policies that adapt to the immediate visual and kinematic state of the surgical field rather than relying solely on preoperative models.

From a control perspective, developing policies for autonomous tumor cutting sits at the intersection of robotics, computer vision, and control theory. Unlike rigid industrial manipulation, the surgical environment is unstructured and highly deformable: tissue responses depend nonlinearly on tool trajectory, speed, contact force, and interactions with neighboring organs. Small changes in tool pressure or approach angle can significantly alter the deformation field, occluding the target or exposing vulnerable structures. While deep Reinforcement Learning (RL) has been successful in complex contact-rich tasks in simulation, its trial–error nature is ethically impermissible on human patients and even problematic on high-fidelity phantoms.

Imitation Learning (IL), particularly Behavioral Cloning (BC), offers a pragmatic alternative by treating surgeon teleoperation as expert demonstrations and learning a mapping from observations (vision and kinematics) to actions. This paradigm can implicitly encode safety constraints, feasible speeds, and surgeon "style" without requiring explicit modeling of tissue mechanics. However, IL in this domain faces two fundamental limitations, well-documented in the general IL literature and increasingly recognized in surgical contexts.

- Covariate shift: Small errors in the learned policy accumulate over time, driving the system into states not represented in the expert data. Because the policy has never been shown how to recover from these off-distribution states, errors compound and can result in large deviations from the intended path—unacceptable when cutting near vital vessels or collecting a margin around a tumor.

- Domain sensitivity and limited generalization: Most surgical IL studies train and evaluate on a single phantom, simulator, or constrained ex vivo setup. In clinical practice, however, there is substantial inter-patient variability in organ shape, tumor size and location, surface texture, and surrounding fat distribution. On top of this, visual appearance shifts with lighting, smoke, blood, and camera pose. Empirically, policies trained on a specific kidney phantom or simulation often fail catastrophically when confronted with geometrically different phantoms, even if the high-level semantics ("cut around the tumor and remove it") are unchanged.

Early efforts in surgical automation relied heavily on kinematic primitives and precomputed trajectories derived from preoperative imaging or fiducial markers. Vision-based closed-loop systems have been demonstrated primarily for suturing tasks, often using near-infrared markers or structured light to track tissue and needle trajectories. STAR's autonomous anastomosis is a prominent example of this line of work. science.org More recent studies have extended autonomy to debridement, irrigation/suction, and soft-tissue retraction, as in Pore et al.'s GAIL-based soft-tissue retraction policy and Ou Tavakoli's RL-based internal tissue point manipulation, but these approaches frequently rely on synthetic phantoms, fiducials, or highly constrained views.

In all cases, explicit tracking often becomes a bottleneck when visual occlusions (smoke, blood, instruments) arise, which is precisely when robust autonomy would be most valuable.

Within RAPN specifically, autonomous tumor cutting remains largely unexplored. The recent work of Furnari et al. on sequence-based imitation learning for kidney tumor removal

is a notable exception: a Vision Transformer (ViT) plus LSTM model is trained on RGB-D video and kinematics to predict future 6-DoF poses of two tools (cautery and grasper) in virtual and physical kidney–tumor environments. iris.unimore.it While this study provides promising evidence that end-to-end IL can capture multi-step tumor cutting behavior, it also highlights key limitations: restricted geometric diversity, limited deformation modeling, and sensitivity to domain shift between virtual and physical setups.

In summary, autonomous tumor cutting in RAPN exemplifies the hardest open problems in surgical robotics: deformable scene understanding, domain-generalizable control, and safe recovery from off-nominal states, all under strict real-time and safety constraints

2.1.1 Deep Imitation Learning in Robotics

The broader robotics community has significantly advanced beyond simple Behavioral Cloning towards deep sequence modeling and large-scale imitation learning, providing methodological tools that are only partially exploited in surgical robotics.

From an algorithmic viewpoint, surveys by Osa et al. and others formalize IL as a spectrum from BC to interactive IL (e.g., DAgger) and adversarial IL (e.g., GAIL and its variants).

These works emphasize that covariate shift and dataset bias are structural properties of BC in long-horizon tasks, motivating strategies that either (i) actively collect data in the states visited by the learner, or (ii) learn implicit rewards that encourage expert-like behavior. GAIL-style methods, such as the soft-tissue retraction framework of Pore et al., show that adversarial IL can improve generalization and sample efficiency compared to pure RL, particularly when simulation is available and real data are scarce.

On the representation side, Transformers have become the dominant architecture for sequential decision making in manipulation. BC-Z (Jang et al.) demonstrated that a single Transformer-based visuomotor policy, trained on more than 100 real-world manipulation tasks, can generalize zero-shot to novel tasks when appropriately conditioned on language or video goals.

More recently, the Open X-Embodiment collaboration showed that large-capacity "robot transformers" (RT-X) can benefit from over one million trajectories collected from 22 distinct robot embodiments, enabling policies that transfer across robots and tasks [17].

These results suggest that scaling model capacity and data diversity systematically improves robustness and generalization in IL, mirroring trends in language and vision.

In the surgical domain, SRT and SRT-H transplant this Transformer paradigm directly into robot-assisted minimally invasive surgery. SRT learns low-level skills such as tissue lifting, needle handling, and knot tying on the da Vinci Research Kit (dVRK) from demonstration, while SRT-H introduces a language-conditioned high-level planner that sequences these skills to perform ex vivo cholecystectomy.

These works provide compelling evidence that Transformer-based IL is well suited to surgical applications, where long-range temporal context and multimodal inputs (vision, kinematics, task phase) are crucial.

Parallel to Transformer policies, Diffusion Policies have emerged as a powerful approach for robot control in general manipulation. Chi et al. treat policy inference as a conditional denoising diffusion process over action or trajectory sequences, showing substantial performance gains across 12–15 diverse manipulation tasks compared to conventional IL and RL baselines.

Diffusion policies excel at modeling multi-modal action distributions, which is highly relevant for tasks where multiple distinct strategies are valid. Early surgical applications—such as diffusion-based suturing and the Diffusion Stabilizer Policy—suggest that these benefits transfer to bi-manual, contact-rich surgical skills as well.

However, diffusion policies come with important trade-offs relative to standard Transformers:

- They require multiple denoising steps per action, increasing inference latency and complicating deployment in high-frequency control loops typical of robotic surgery.

- They are data-hungry and architecturally more complex, which is problematic in surgical settings where high-quality demonstrations are limited.

- Their inherent stochasticity complicates verification and validation in safety-critical scenarios.

In contrast, causal Transformer policies offer a single forward pass per control step, deterministic behavior (unless randomness is explicitly injected), and a simpler training pipeline, making them attractive for real-time surgical applications with tight latency and safety constraints. Our work therefore adopts a Transformer-based IL approach rather than a diffusion policy, explicitly prioritizing real-time feasibility, interpretability, and robustness in the small-data regime that characterizes current RAPN datasets, while still leveraging the sequence modeling strengths demonstrated in SRT, BC-Z, and RT-X.

2.1.2 Representation Learning for Control

Handling high-dimensional surgical video while maintaining precise control over SE(3) tool motion is fundamentally a representation learning problem. In general robotics, self-supervised visual pretraining (e.g., contrastive learning, masked autoencoding) has become standard practice to initialize encoders on large unlabeled datasets before fine-tuning on specific tasks. In surgical computer vision, similar strategies have been used for instrument segmentation, phase recognition, and workflow analysis, and domain-adaptation techniques such as the teacher–student framework of Sahu et al. have been proposed to transfer segmentation models from synthetic to real endoscopic data.

Yet, for precise manipulation, the trajectory itself carries rich semantic and geometric information: surgical tool motion is highly structured, approximately smooth, and constrained by kinematics, anatomy, and surgical intent. Recent work in sequence-based IL for tumor removal by Furnari et al. makes this explicit by using an LSTM to model temporal dependencies in the 6-DoF pose sequences of two tools during kidney tumor cutting and removal, showing that the temporal structure of trajectories is crucial for success.

Despite this, there is still very limited exploration of trajectory-centric pretraining in surgical robotics. Most IL methods treat action prediction as an end-to-end mapping from raw pixels to joint or Cartesian commands, without explicitly encoding the geometric and dynamical regularities of surgical motion. In particular:

1. The SE(3) structure of tool motion is rarely exploited in the representation (e.g., through equivariant models or motion-specific augmentations).

2. Kinematic sequences are generally used only as supervised training signals, not as a pretraining resource to learn a "motion manifold" independent of vision.

3. Domain adaptation focuses primarily on visual features (e.g., sim–real image translation), rather than on generalizing motion patterns across organ geometries and phantoms.

This gap is particularly problematic for tumor cutting in RAPN, where the geometric variability of kidneys and tumors is high, but the underlying surgical motion primitives (approach trajectories, margin-respecting cuts, safe withdrawal) share strong structural regularities that could be learned from kinematics alone and then combined with vision.

2.1.3 Gap Analysis and Contributions

Building on insights from general robotics IL, surgical autonomy, and domain adaptation, we identify three gaps that currently limit progress toward clinically meaningful autonomous tumor cutting:

1. Lack of geometric diversity in benchmarks. Existing benchmarks and datasets for autonomous surgical manipulation predominantly rely on repeated trials on identical or near-identical phantoms, or on single synthetic environments in simulation. DESK [18], for example, targets cross-robot gesture transfer in peg-transfer tasks rather than organ-level cutting Furnari et al. [19] introduce a valuable dataset for kidney tumor removal, but geometric diversity in phantom anatomy and tumor morphology is still limited, and domain shifts between virtual and physical setups are only partially explored.There is currently no standard benchmark that systematically varies kidney geometry, tumor shape and location, and superficial appearance to rigorously assess whether an autonomous policy truly generalizes across anatomies, rather than overfitting to a single phantom.

2. Under-utilization of kinematic structure. Conversely, general IL literature shows that modeling trajectory structure explicitly, through segmentations, options, or motion primitives, can greatly improve robustness and data efficiency [20]. In autonomous tumor cutting, where the shape of the desired path is tightly constrained by safety margins and tissue deformation, ignoring the strong priors embedded in surgical tool motion is a missed opportunity. Moreover, few works incorporate explicit deformation estimates (e.g., from stereo depth or learned organ models) into the state, even though deformation fundamentally affects the safe cutting trajectory.

3. Insufficient domain adaptation in control policies. While domain adaptation has been studied extensively for surgical perception (e.g., sim–real instrument segmentation via teacher–student learning) and for high-level shared control frameworks, it is rarely integrated directly into low-level control policies for cutting tasks [21–23]. Pore et al. show that GAIL can transfer soft-tissue retraction policies from simulation to synthetic phantoms, and Zhang et al. use sim-to-real adaptation with DMPs for shared control, but similar techniques have not been systematically applied to organ-level resection policies [24]. As a result, most autonomous cutting policies are evaluated in the same domain in which they were trained, leaving their robustness to real-world domain shifts largely untested.

2.1.4 Structure of the Chapter

Given these challenges and motivations, this chapter presents a two-stage approach to imitation learning for surgical robotics. The first part focuses on learning in a simulated environment, where high-fidelity virtual organs and tools enable safe, large-scale data generation. This environment provides a platform to explore and validate the core learning architecture under controlled and repeatable conditions, forming a bridge between conceptual design and practical implementation.

The second part transitions to real-world validation using physical phantoms constructed from silicone and derived from patient CT data. This section addresses the simulation-to-reality gap, demonstrating how learned policies can be executed by a real robot interacting with deformable materials. The validation on phantoms provides critical insights into robustness, generalization, and the practical challenges of implementing imitation-based autonomy in physical surgical scenarios.

Together, these two stages form the foundation of learning-based surgical autonomy: simulation enables the development and refinement of algorithms, while physical validation confirms their feasibility in realistic operative conditions. The chapter ultimately demonstrates how a

robot can learn both actions and decisions from demonstration, forming a key component of the broader framework of intelligent and context-aware surgical autonomy.

2.2 Stage I: Learning from Demonstrations in a Virtual Environment

Significant strides have been made in the field of autonomous surgical operations, with researchers exploring various approaches to enhance the capabilities of surgical robots and improve patient outcomes [25] [26] [27] [28] [29]. Though, integrating autonomous procedures into surgical practice remains a formidable challenge [1] [30]. Unlike other domains where supervised learning approaches have seen widespread adoption, such as image recognition and natural language processing, its application in surgical tasks is limited. This limitation primarily stems from the scarcity of high-quality datasets, either simulated or real-world, necessary for training and validating machine learning models for surgical applications. In surgical robotics, where precision, safety, and reliability are of paramount importance, supervised learning poses unique challenges. Unlike tasks with well-defined goals and abundant training data, surgical procedures involve intricate spatial-temporal interactions and anatomical complexities. Training models solely on labeled data often fails to capture the nuances of surgical tasks, leading to sub-optimal performance and limited generalization capabilities. The majority of robot surgical datasets are made for phase detection of procedures [31] [32] or object (such as organs or tools) detection and segmentation [33] [34] [35]. These kinds of dataset lack of either kinematic data or execution of task in environments with different morphologies, therefore generalization across different scenarios would be difficult.
The JIGSAWS dataset [36], is probably the most known dataset in surgical robotics. This dataset includes 76-dimensional kinematic data along with the video data for 101 trials of three elementary surgical tasks (suturing, knot-tying, and needle-passing) performed by six surgeons using the da Vinci Research Kit (dVRK), but it does not execute tasks on different scenarios. Similar considerations can be asserted for other works which use dVRK [37]. To address these challenges, this work proposes a model to learn task execution in unseen environments. This necessitates learning a policy that dictates the agent's behavior based on its observed state. We achieve this by leveraging imitation learning [38], where agents learn by imitating expert demonstrations. This approach is particularly valuable in surgical robotics, where purely vision-based or geometric approaches for policy generalization across diverse environments are difficult, if not impossible.
By leveraging expert demonstrations, it empowers agents to learn complex tasks by mimicking skilled human behavior.
In domains such as robotic manipulation, imitation learning has proven particularly effective for tasks like pick and place operations [39] [40]. In pick and place tasks, the focus is often on learning the optimal picking and placing poses rather than the entire trajectory. Another prominent application of imitation learning is in the field of autonomous driving [41]. Here, agents learn to navigate complex road environments by imitating the driving behavior of human experts. Unlike continuous control tasks, actions in autonomous driving are often discrete, such as turning left or right, braking, or accelerating.
Despite its successes in other domains, while imitation learning has the potential to teach agents to perform surgical tasks based on observed expert behavior, its application in surgery has been limited, primarily due to the scarcity of high-quality datasets and the unique challenges posed by surgical environments. Some datasets in the field of surgical tasks do actually exist but they are made for segmentation and phase recognition. In the context of surgical robotics, imitation learning offers a promising avenue for teaching agents surgical procedures based on observed expert actions and circumvents the need for explicit programming or manual specification of task rules, allowing agents to acquire skills through demonstration and imitation, Though proposed

models lack generalization across different scenarios [42] or use different approaches and very simple tasks [43] [44] [45].

The main contributions of this section are:

- We address the scarcity of high-quality data in surgical robotics by introducing a novel dataset of synthetic kidney-tumor environments. This dataset is generated by leveraging advanced 3D reconstruction techniques, based on real patients' CT scans to create accurate kidney and tumor models. These models are then imported into a simulation environment, enabling the simulation of surgical tasks like tumor removal. This approach provides a realistic and diverse set of environments for training and testing autonomous surgical systems.

- We propose a novel neural network architecture specifically designed for imitation learning in surgical robotics. This architecture combines the strengths of Vision Transformers (ViT) and Long Short-Term Memory (LSTM) networks. The ViT component effectively extracts spatial features from visual data, crucial for tasks like tumor segmentation in the kidney environment. The LSTM component excels at capturing temporal dependencies, essential for learning the sequential nature of surgical procedures. By combining these capabilities, our model learns to execute tumor removal tasks in new kidney-tumor environments. Leveraging imitation learning techniques, the model learns from expert demonstrations in similar environments and adapts this knowledge to perform tasks autonomously in previously unseen scenarios. This demonstrates the efficacy of policy learning for autonomous surgical robotics.

2.2.1 Dataset Creation

Generation of 3D Environments from Patient CT Scans

Integrating autonomous procedures into surgical practice remains a formidable challenge, primarily due to the scarcity of high-quality datasets and the complexity of modeling dynamic tissue behavior. The first stage of this research addressed this issue by creating a comprehensive synthetic dataset of surgical procedures in a simulated environment, allowing controlled exploration of learning architectures and domain variability. The objective was to teach a model how to execute tumor removal in previously unseen environments by imitating expert demonstrations and generalizing across differences in shape, size, and position of the tumor. Acquiring a dataset that accurately represents surgical scenarios poses significant challenges. In this section, we detail the creation of a dataset for imitation learning focused on tumor removal from the kidney, leveraging advanced simulation techniques and medical imaging technology. While our study focused on the specific task of kidney–tumor removal, the potential applications of imitation learning in surgical robotics extend far beyond this single procedure. Its versatility and adaptability make this work suitable for a wide range of surgical tasks across different anatomical structures and organs. We also envision expanding this research to encompass a broader array of surgical procedures. To enhance generalization, a dataset was created across multiple simulated environments varying in:

- Tumor size, position, and orientation.

- Lighting, color, and texture of organs.

- Starting pose and orientation of the surgical tool.

This variability was intentionally introduced to promote **domain adaptation**, pushing the model to learn underlying surgical strategies rather than memorizing visual patterns. Each

demonstration consisted of synchronized RGB images, depth maps, and Cartesian trajectories of the tool tip.

We utilized the SOFA (Simulation Open Framework Architecture) software due to its capability to handle real-time deformation of soft objects, akin to human organs. The deformable nature of organs, such as the kidney, presents a complex challenge for surgical simulations. By applying empirically determined deformation parameters, including the elasticity modulus, it is possible to realistically simulate the behavior of soft tissues during surgical manipulation.

By default, the SOFA framework uses the *Young's modulus* (E) and *Poisson's coefficient* (ν) as deformation parameters. Through empirical experimental tests [46], it has been demonstrated that the elasticity modules of the kidneys are $E = 180.32\ kPa$ and $G = 95.64\ kPa$ (tangential elasticity modulus), which can be related by the following formula:

$$G = \frac{E}{2(1+\nu)} \tag{2.1}$$

From this relation we obtained $\nu = -0.057$. The mechanical characteristics of renal tumors, as highlighted in the research by Levillain *et al.* [47], may vary considerably based on the tumor type, but they are generally softer than renal tissue. The following values were empirically identified as those that ensure biomechanical behavior as realistic as possible: $E = 5\ kPa$ and $\nu = -0.8$.

The goal is to faithfully replicate the physical stresses that occur during actual surgery, including gravitational forces, tissue deformations, and collisions between the instruments and the organs. The procedure considered as a use case involves the use of one instrument to remove the tumor from the surrounding tissue area and another tool to grasp and remove it. Every phase of this procedure was reproduced with attention to physical realism, ensuring that instruments behave consistently and that the tumor is removed precisely. In this way, it is possible to accurately emulate the removal of a renal tumor through robotic surgery.

A fundamental aspect is the generation of a complete dataset that includes realistic variations in tumor and organ conditions. This dataset is used to train an imitation learning model, which learns from video recordings of demonstrations and tracking of kinematic data to autonomously perform the surgical procedure on new scenarios.

Video Demonstrations Execution

to capture surgical demonstrations for the imitation learning dataset, we employed a manual execution approach conducted by a single engineer, as our primary focus was on testing the model across different environments rather than evaluating variability across different experts in the same environment. This approach allowed us to assess the model's adaptability and robustness in diverse scenarios. The demonstrations were made using specialized surgical tools within a simulated environment. The task of tumor removal from kidneys was executed using two distinct instruments: a cauter for the cutting phase and a grasper for grasping and removing the tumor post-incision. These instruments were controlled manually via a console controller, allowing for precise manipulation within the simulation environment. The integrated RGB-D camera is used as stereoscope to capture videos of the demonstrations. The cauter, employed for the cutting phase, simulated the use of electro-cautery to incise tissue. This instrument was responsible for executing precise incisions along predetermined trajectories within the kidney-tumor environment. Conversely, the grasper functioned as a versatile tool for grasping and removing the dissected tumor fragments from the surgical site. Together, these tools enabled the execution of a complete tumor removal procedure within the simulation. The execution of surgical demonstrations was facilitated by a modified version of the tissue dissection application developed within the SOFA ENV framework, utilizing LapGym[27]. This application was tailored to our specific use case, providing a specialized environment for executing kidney-tumor removal procedures. By leveraging the capabilities of LapGym, a laparoscopic surgery simulation module

within SOFA ENV, we were able to recreate realistic surgical scenarios while allowing manual control of surgical instruments. The cutting procedure starts from the left side of the tumor and proceeds follows clockwise along the visible perimeter. Once the cutting edge has completed more than half of its trajectory and has passed the gripping tool, positioning born on the right side, the latter approaches the tumor to grasp it and establish one secure grip, holding it in place for the final stage of the procedure. Once removed all contact points the tumor is completely separated from the kidney and must be removed from the surrounding healthy tissue: the task can be considered finished successfully when the tumor was positioned at a safe distance from it to the organ. Some of these steps are shown in Figure 2. It is very important to carry out all simulations always respecting the same guideline to ensure a certain consistency between the dataset demonstrations: not only is it essential to carry out the task correctly but it is equally crucial to obtain a homogeneous dataset so that, during the training phase, the learning agent is able to recognize visual and sequential patterns in the action preformed across different environments.

It is important to clarify that the demonstrations used for dataset creation in this chapter were performed by engineers rather than clinical surgeons. While expert surgeon demonstrations would represent the ideal source of data, this choice was motivated by both methodological and practical considerations intrinsic to early-stage autonomous surgical robotics research.

The primary objective of this work is to enable the robot to learn kinematic behaviors from visual input, including tool trajectories, orientations, and temporal coordination, rather than high-level surgical decision-making. The considered tasks were therefore designed to focus on geometric consistency, motion smoothness, and reproducibility of expert-like trajectories under controlled conditions. These aspects can be reliably demonstrated by trained engineers following clinically inspired protocols and anatomical constraints derived from surgical practice and literature.

Also this strategy enables rapid iteration, large-scale data collection, and safe evaluation of learning algorithms without imposing unnecessary demands on clinical personnel or exposing them to immature robotic systems.

Nevertheless, the involvement of expert surgeons remains a crucial step toward clinical translation. Future work will focus on collecting demonstrations from surgeons and leveraging real surgical video data to further refine the learned policies. The framework developed in this chapter is explicitly designed to support this transition, providing a robust foundation upon which surgeon-driven learning and validation can be performed once the system reaches sufficient maturity and safety.

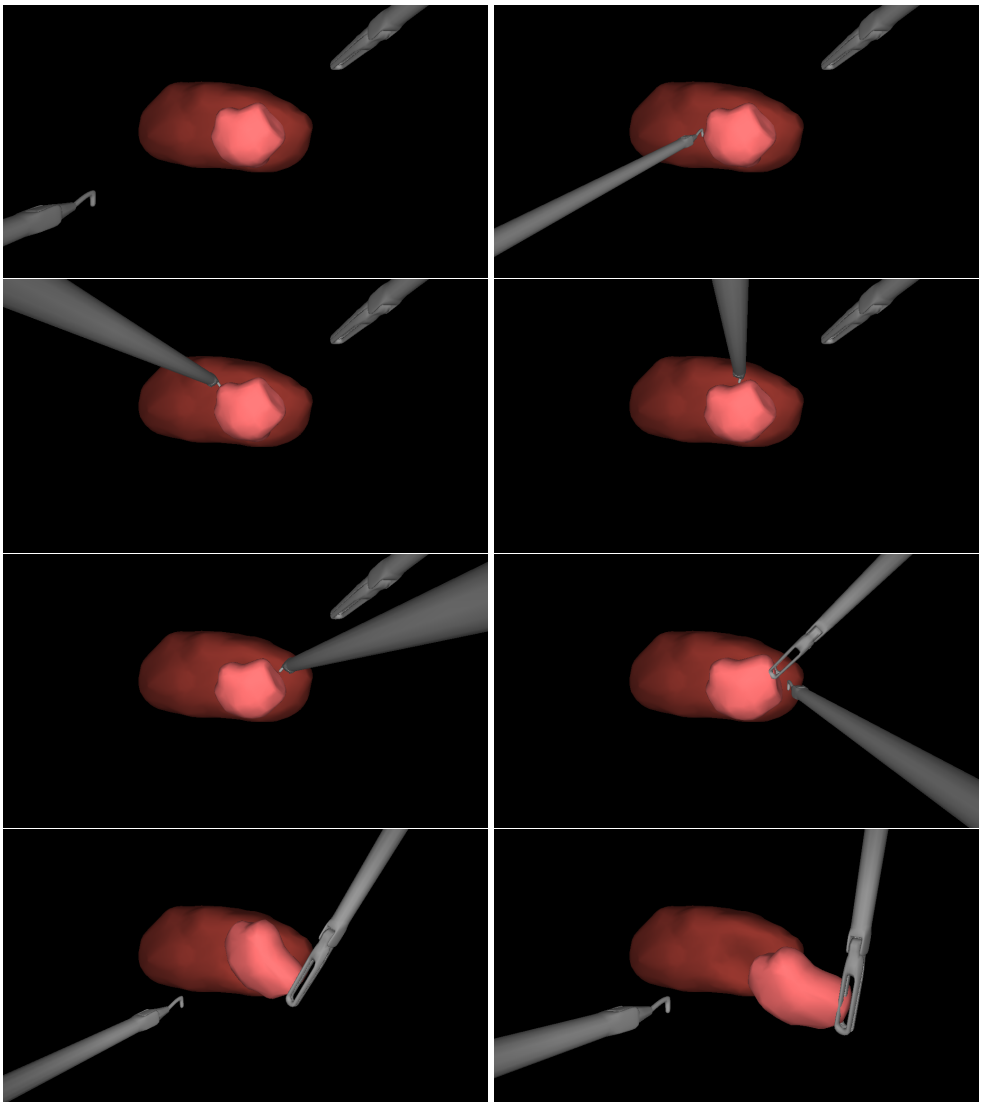


Figure 2.1: Some steps of the execution of tumor cutting and removal from kidney executed by the expert. The cauter (tool on the left) cuts the tumor and the grasper (tool on the right) take the tumor out of the kidney

.

Dataset Composition

The dataset for our imitation learning project is composed of multi-modal data captured during the execution of surgical demonstrations within the simulated kidney-tumor environments. Each demonstration is recorded using an RGB-D camera, providing both visual and depth information essential for understanding the surgical scene. The primary component of the dataset consists of videos captured through RGB-D cameras, with each demonstration recorded as a sequence of frames. These videos offer a detailed visual representation of the surgical procedure, allowing for the observation of tool manipulation and tissue interaction throughout the task execution. In addition to RGB video frames, depth information is also captured for every frame of each video using the same RGB-D camera. This depth data provides crucial spatial context, allowing for accurate estimation of the 3D geometry of the surgical scene. Depth information is invaluable for understanding the relative positions of surgical instruments, tissues, and anatomical structures within the environment. Alongside video and depth data, the dataset includes precise 3D pose information for both the cauter and the grasper instruments. The pose data comprises the XYZ position and unit quaternion for orientation of the tips of both tools at each time step of the

trajectory. These tool poses serve as ground truth data for our imitation learning model, guiding the prediction of tool actions in subsequent time steps. Additionally, two more variables, one for each tool, are provided, describing the sate of the tools: for the cauter, a boolean variable is considered with value of 1 if the tool is cutting, 0 otherwise; for the gransper the variable considered indicates the opening angle of the tool. In conclusion, the creation of the dataset described here in was necessitated by the absence of suitable resources for learning surgical tasks in the intended manner. Despite the growing interest in surgical automation and imitation learning, existing datasets fail to adequately address the specific requirements of our task. Real-world surgical datasets often lack the level of detail and precision required for effective imitation learning. To sum up, each demonstration included:

- RGB and depth frames from the stereo camera;
- 3D positions and orientations (XYZ + quaternion) of both tools at each time step;
- Tool state indicators (cutting activity and grasper opening angle).

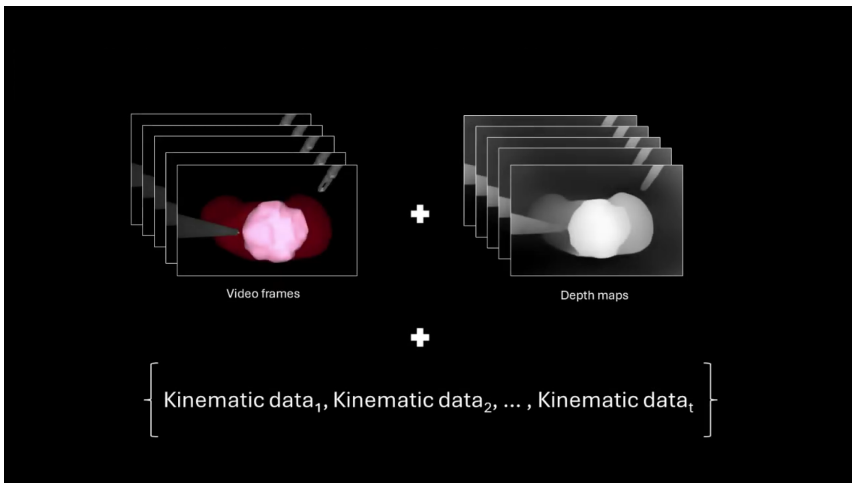


Figure 2.2: Composition of a single element of the dataset: RGB-D images plus kinematic data of the two tools executing the operation

.

This multimodal dataset captures both spatial and temporal characteristics of the task, forming the foundation for policy learning. By encompassing multiple synthetic organ morphologies, it encourages generalization and robustness across domains, a crucial aspect for transfer to real-world scenarios.
Obtaining annotated data from actual surgical procedures is challenging due to ethical and logistical constraints. Additionally, real surgical environments may exhibit variability that is difficult to control and replicate consistently. Of course a challenge for the future will be to have a good-quality dataset with real-world data. On the other side, while simulation platforms would offer a controlled and safe environment and a convenient, albeit preliminary, way for data collection, there are very few platforms where surgical environments can be simulated and existing datasets within these platforms often nuance details necessary for accurate imitation learning, such as realistic tissue deformation, instrument-tissue interaction dynamics and kinematics data of the tools exploited. Furthermore, the specific task of tumor removal from kidneys presents unique challenges that necessitate the creation of a tailored dataset.

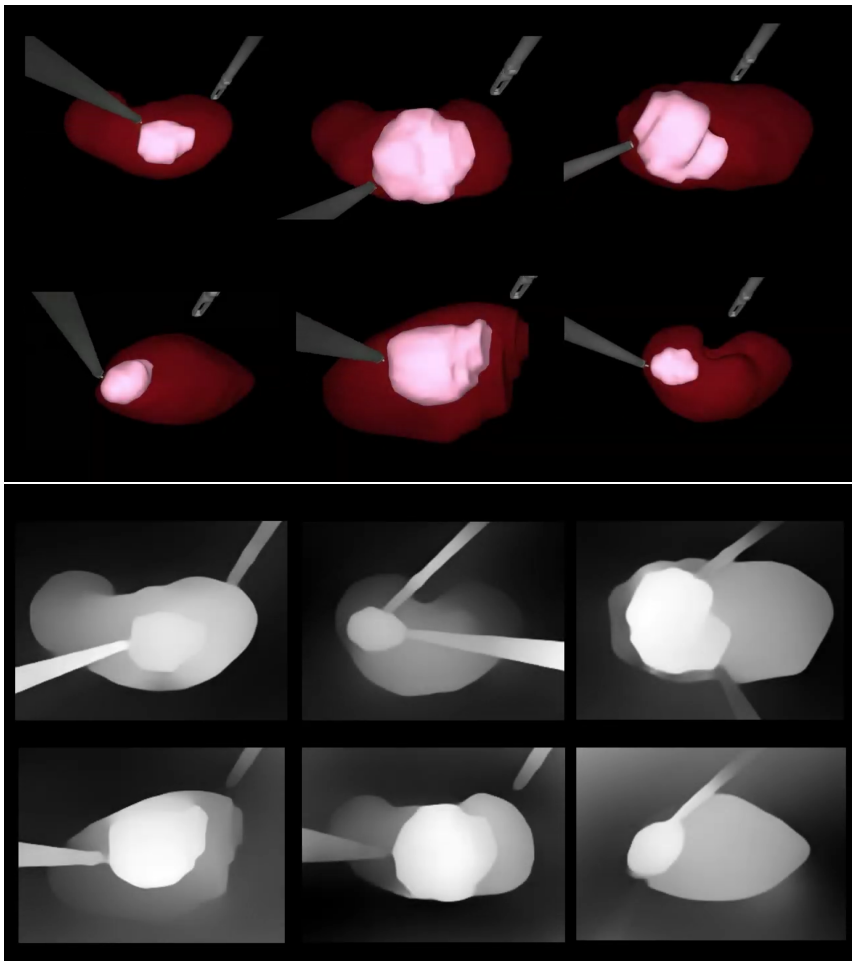


Figure 2.3: Different domains in datset composition

.

2.2.2 Imitation Learning Model

Problem Formulation

At its core, imitation learning seeks to train an agent to replicate expert actions based on observed demonstrations, without requiring explicit task specification or reward design. One common formulation of imitation learning is *behavioral cloning from observation* [48], where an agent learns to imitate expert behavior by directly mapping observed states to corresponding actions. In this paradigm, the agent aims to replicate the demonstrated behavior without explicitly modeling the underlying task dynamics. Instead, it learns a mapping from observed states to actions through supervised learning techniques.

Central to the success of imitation learning is the notion of *policy learning*. A policy represents the strategy or decision-making function employed by an agent to select actions based on its observations of the environment. In this work, we employ behavioral cloning to train an agent for performing surgical tasks, such as tumor removal from kidneys. At each step of task execution, the agent must decide on an action based on the current state of the environment and past states. This involves finding a policy—a function of the current visual observation and the state of the surgical tools (e.g., kinematic data)—that dictates the agent's actions to achieve the desired surgical outcome, namely the state the tools must assume in the next time step.

We can formulate the problem as follows. Let s_t denote the current state of the environment at time step t, and a_t represent the action taken by the agent in response to s_t. The agent learns a mapping from states to actions, represented by a policy $\pi(s)$. This mapping is learned

from a dataset of expert demonstrations $\{(s_1, a_1), (s_2, a_2), \ldots, (s_N, a_N)\}$, where N is the number of demonstrations. The goal of policy learning is to find an optimal policy π^* that minimizes the discrepancy between the agent's actions and the expert actions. In the case of supervised learning, this discrepancy is typically measured using a loss function $L(\pi(s), a)$:

$$\pi^* = \arg\min_{\pi} \sum_{i=1}^{N} L\big(\pi(s_i), a_i\big) \tag{2.2}$$

where s_i and a_i are the states and actions from the expert demonstrations, respectively.

In our context, the agent's action at each time step is determined by a policy $\pi(s_t, \theta)$, parameterized by θ. The policy is learned to mimic the expert behavior observed in the demonstrations, where θ is optimized to minimize the loss between the agent's actions and the expert actions, as follows:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} L\big(\pi(s_i, \theta), a_i\big) \tag{2.3}$$

where N is the number of demonstrations. The state s_t is composed of the observation and the robot tools' state (their kinematic data), denoted as $s_t = (o_t, tools_t)$. Additionally, the parameter θ depends on past states, reflecting the sequential nature of policy learning in imitation learning tasks.

Let s_t denote the state of the environment at time step t, consisting of the observation and the predicted action at time t. Let w represent the size of the temporal window of past states, including both past observations and past predicted actions. The action at time step $t + 1$, denoted as a_{t+1}, is obtained from the application of the policy function $\pi(s_t, \theta)$ to the current state s_t, along with the window w of past states $\{s_{t-w+1}, s_{t-w+2}, \ldots, s_t\}$:

$$a_{t+1} = \pi(s_t, s_{t-1}, \ldots, s_{t-w}) \tag{2.4}$$

where $\pi(s_t)$ is the policy function that takes as input the current state s_t and predicts the action a_{t+1} to be executed at the next time step.

Each state s_t in the sequence window is composed of both past observations and past predicted actions:

$$s_t = (o_t, a_t) \tag{2.5}$$

where o_t represents the observation at time step t, and a_t represents the predicted action at the same time step. This sequential formulation captures both spatial and temporal dependencies in surgical trajectories, allowing the model to learn realistic motion dynamics from expert demonstrations.

Proposed Architecture

The imitation learning model employed in our study leverages a hybrid architecture comprising a Vision Transformer[29], pretrained on ImageNet, and a Long Short-Term Memory (LSTM) network as shown in Figure 3. This architecture is designed to effectively handle the multi-modal nature of the dataset, integrating both visual information from RGB-D frames and temporal sequences of tool poses. The model takes, as input, sequences of RGB-D frames along with past predictions iteratively to predict the actions that the surgical tools should take in the next time step, i.e. the poses every tools has to assume in the next step. By iteratively refining its predictions based on past observations and ground truth tool poses, the model learns to mimic the expert demonstrations and perform surgical tasks autonomously in new kidney-tumor environments. The Vision Transformer serves as the backbone of our model, specifically configured with 16 patches and a resolution of 224 pixels, responsible for processing the visual information

extracted from RGB-D frames. By employing self-attention mechanisms, the ViT can capture spatial relationships and long-range dependencies within the video data, enabling effective feature extraction from each frame. This allows the model to encode rich visual representations of the surgical scene. In our model, the LSTM network is a sequence-to-one architecture, where it takes a sequence of input features and predicts a single output value. It plays a crucial role in processing sequences of features extracted by the ViT: it is responsible for capturing temporal dependencies and learning the dynamics of the surgical task over time. We consider the trajectories of surgical tools as sequences of steps, where the action at time step t is influenced by both the preceding steps and the current state of the environment. This approach allows us to effectively model the long-term dynamic of the trajectories in surgical operations. We opted for an LSTM network instead of a traditional deep network because LSTMs are specifically designed to handle sequential data, such as the trajectories in our study. The LSTM's ability to retain information over time and capture dependencies between time steps allows it to model the sequential nature of surgical operations more effectively. This approach resulted in better performance and more accurate predictions compared to using a standard deep network. Moreover, an attention mechanism is incorporated within the LSTM architecture to enhance the model's ability to focus on relevant parts of the input sequence. The LSTM layer processes the input sequence, which consists of features extracted by the ViT at time step t and the previous w features, along with the corresponding predicted actions of the robot tools. This sequence is fed into the LSTM, allowing the model to capture long-term dependencies and temporal dynamics. Following the LSTM layer, an attention mechanism is employed to dynamically weigh the importance of each input feature in the sequence. The attention mechanism computes attention weights using a linear layer applied to the LSTMoutputs. These weights reflect the relevance of each feature in the sequence, allowing the model to focus more on informative features while attenuating the impact of less relevant ones. The attention weights computed by the linear layer are used to compute a weighted sum of the LSTM outputs, yielding an attended input representation. This attended input is then passed through another linear layer to produce the final output, which represents the action the robot tools are expected to take at the next time step. The integration of the attention mechanism within the LSTM architecture enhances the model's performance by enabling it to selectively focus on relevant information in the input sequence, thereby improving its predictive capabilities.

Training of the Model

The model predicts the actions of the surgical tools through a sequence of processing steps, which unfold as follows. The RGB-D frame at time t is divided into patches and passed through the Vision Transformer (ViT) and a multilayer perceptron (MLP) to extract high-dimensional features that capture spatial and temporal information.

The features extracted at time t are combined with features from previous time steps $t-1$ to $t-w$, where w is the window size representing the number of past states considered in the input sequence. Additionally, past predicted actions from time t to $t-w$ are included in this fusion process. This temporal fusion mechanism allows the model to leverage historical information and past predictions to perform context-aware action prediction.

The temporally fused features are then passed into an attention-based LSTM block, which comprises three main components: an LSTM layer, an attention mechanism, and an MLP. The LSTM layer processes the temporally fused features, capturing temporal dependencies and dynamics across the input sequence. The attention mechanism computes attention weights using a linear layer applied to the LSTM outputs. These weights reflect the relevance of each feature in the sequence, allowing the model to focus more on informative features while attenuating the influence of less relevant ones. The output of the attention mechanism is subsequently passed through the MLP to produce the final prediction.

The MLP outputs a 16-element vector comprising the x, y, z positions and unit quaternions

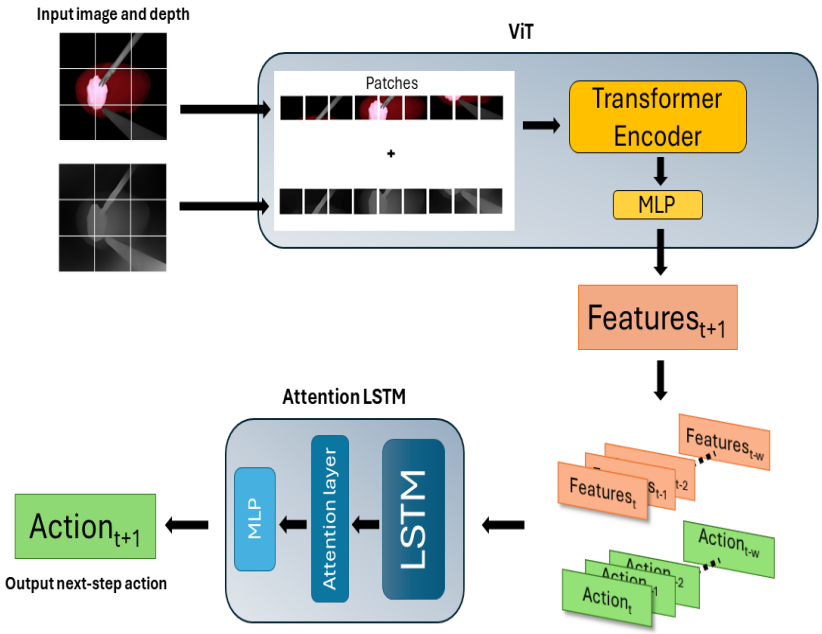


Figure 2.4: Cycle to predict an action for a single step of the trajectory

.

of the two surgical tools, along with the cutting or not-cutting Boolean variable for the cauter and the opening angle of the grasper. This process is repeated for each point in the trajectory and for every trajectory in the dataset (i.e., each video in the dataset), allowing the model to learn and generalize across diverse surgical scenarios.

For training, the Adam optimizer was employed with an initial learning rate of 10^{-5}, which was reduced to 10^{-6} if the loss did not improve for five consecutive epochs. This adaptive learning rate schedule helps the model converge effectively and prevents it from becoming trapped in local minima. A window parameter of $w = 10$ was empirically determined to provide a good balance between capturing temporal dependencies and maintaining computational efficiency. This choice allows the model to consider a sufficient history of past states and actions while avoiding excessive computational overhead.

The model was trained for 50 epochs. Experimental results showed that training beyond this point did not yield significant improvements in performance, indicating convergence to a stable solution.

A custom loss function, referred to as the *Temporal Loss*, was introduced to address the challenge of error accumulation over long prediction horizons. This loss function penalizes errors more heavily as predicted actions deviate further from the ground truth over time. By assigning higher weights to errors in later time steps, the Temporal Loss encourages the model to prioritize accurate predictions for actions in the immediate future while still considering the global consistency of the surgical trajectory. This approach mitigates the accumulation of prediction drift often encountered in sequential tasks, where uncertainties and compounding errors tend to increase as the task progresses.

The dataset was divided into training and testing sets using an 85/15 ratio. This split ensures that the model is trained on a sufficiently large portion of the data while reserving a separate subset for evaluating generalization performance on unseen examples.

2.2.3 Results in the Virtual Environment

In our study, we utilized two tools to predict trajectories: a grasper and a cauter. The grasper presented a relatively straightforward challenge, as both its trajectory and the associated task were simple and easy to model. For this reason, the success rate was 100% for the grasping part of the task. However, the cauter posed a more significant challenge due to the complexity of its trajectory and the precision required for the task. Predicting accurate trajectories for the cauter is crucial, as it directly impacts the effectiveness and safety of the surgical procedure. Consequently, the results section primarily focuses on the performance of the model with the cauter, as it represents the more demanding aspect of our work.

During training, we monitored the Mean Absolute Error (MAE) on positions, achieving an average result of 0.5 cm. This metric quantifies the average discrepancy between the predicted and ground-truth positions of the robot tools, providing insight into the model's accuracy in mimicking expert behavior.

Following the training phase, we evaluated the model's performance on previously unseen environments to assess its generalization capabilities and effectiveness in simulated scenarios. The trained model demonstrated a commendable success rate, correctly executing the surgical task in 70% of new environments (7 out of 10 total tests). This result underscores the model's ability to generalize from training data to unseen test environments and to effectively replicate expert behavior across diverse surgical settings.

In some scenarios, the model generated trajectories that surpassed the quality of the target trajectories observed in expert demonstrations, as shown in Figure 2.7. These generated trajectories exhibited smoother and more continuous motion, effectively avoiding uncertainties, shakings, and corrections typically associated with manual execution. This finding indicates the model's potential to enhance the precision and efficiency of surgical procedures by minimizing unnecessary movements and ensuring consistent tool manipulation.

The results of our study indicate that there are no significant differences in performance between right and left tumors. Additionally, translation of the tumor in space, meaning the position of the tumor relative to the kidney, does not have a substantial impact on the outcomes. This suggests that the model is robust to variations in tumor location and side, maintaining consistent performance regardless of these factors.

Despite its overall success, the model encountered difficulties in environments where the tumor shape and size deviated significantly from the norm. In such cases, characterized by unusually shaped tumors or exceptionally large sizes, the model exhibited reduced performance and struggled to execute the task accurately; that is, the generated trajectory deviated considerably from the target trajectory. As a result, the tumor may not be fully removed from the kidney, or the surgical margins may be excessive, leading to the removal of more healthy tissue than necessary. These outlier scenarios present unique challenges for the model, highlighting the need for further refinement to ensure precise and accurate surgical outcomes, as well as robustness and adaptability in real-world applications.

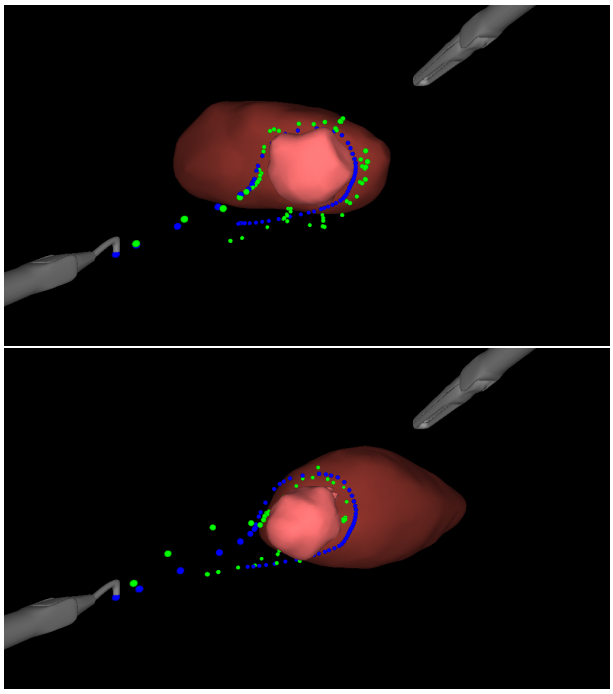


Figure 2.5: Two examples of successful task executed generating the trajectory using our model. In the picture the target trajectory is highlighted in green while the generated trajectory is blue. In the first case we can notice that the generated trajectory is smoother and more accurate than the target trajectory.

.

2.2.4 Discussion and Evolution Toward Real Environments

The project described in this work demonstrates the potential of imitation learning for automating surgical tasks. By leveraging synthetic environments and expert demonstrations, the trained model exhibits promising capabilities in executing complex surgical procedures autonomously.

While direct comparisons with other works are challenging due to differences in tasks, methods, and environments, we can still highlight several key distinctions. Unlike our study, which emphasizes the model's performance across different scenarios and environments, most existing works focus on variations in expert execution within a single environment. Furthermore, the tasks addressed in other studies often differ in complexity and objectives, making it difficult to draw straightforward parallels. Despite these differences, this work contributes uniquely by exploring the adaptability of the model across varying conditions, which is less commonly addressed in the current literature.

In particular, the creation of the dataset and the development of the model represent a significant contribution to the field of surgical robotics and automation. The dataset described in this study fills a critical gap in surgical imitation learning by providing a comprehensive and diverse training resource. Successful imitation learning relies on exposure to a wide range of environments. By creating a dataset encompassing multiple kidney–tumor scenarios, the model learns to adapt to different anatomical configurations and surgical challenges.

While the results are promising, challenges remain in handling outlier scenarios. When the tumor shape or size differs substantially from the norm, the model is not always able to execute the task successfully. This emphasizes the need for continued refinement and adaptation of the model to ensure robust performance in diverse surgical settings. Additionally, one of the primary challenges moving forward is the development of more realistic surgical environments. Current synthetic setups, while useful for training and initial testing, may lack the complexity and variability of real-world surgical conditions. Future efforts should aim to enhance the fidelity

of simulated environments by incorporating more realistic tissue textures, adding other organs to alter the executed trajectories, and including additional visual elements such as blood and fat.

Moreover, the tasks modeled in this phase were simplified and do not fully capture the complexity of real surgical procedures. While such simplification is beneficial for initial experimentation, it does not encompass the complete range of challenges encountered in real-world surgery.

Transitioning from synthetic to real-world environments, both *in silico* and on physical organ phantoms, presents a significant challenge for autonomous surgical systems. Real environments introduce numerous complexities, including variability in patient anatomy, tissue characteristics, intraoperative conditions, and robot control. These factors can lead to deviations in accuracy and stability during task execution. Acquiring high-quality real-world datasets that capture this variability is essential for training models capable of generalizing effectively across diverse surgical scenarios.

While the research presented here has primarily focused on automating specific surgical tasks such as tumor removal from kidneys, there remains a clear need to expand the scope to a broader range of procedures, including suturing, blood suction, and dissection. Addressing such varied procedures will enable a more holistic understanding of how autonomous systems can assist surgeons throughout the full spectrum of operative tasks.

As the field of surgical robotics and automation continues to advance, tackling these future challenges will be critical for the development of robust and versatile autonomous systems. By integrating advances in machine learning, computer vision, and robotics, researchers can overcome these obstacles and pave the way for safer, more efficient, and more accessible surgical interventions.

The experience gained in simulation laid the foundation for real-world deployment. The next stage transferred these insights to a physical setup with a UR5 robot performing tumor removal on phantoms, supported by a refined architecture and trained on real multimodal data.

2.3 Stage II: Learning from Demonstrations on Robotic Phantoms

2.3.1 Introduction and Motivations

A key motivation for this second stage is its direct connection to the first part of the chapter, where imitation learning was developed and validated entirely within a virtual environment. The simulation phase provided an essential foundation: it allowed rapid data collection, large-scale experimentation, and systematic architectural refinement in a controlled setting free from hardware limitations. However, policies trained in simulation alone cannot fully account for the visual complexity, sensor noise, and physical dynamics present in real surgical scenes. The present stage therefore builds on the strengths of the virtual environment while addressing its inherent limitations. By deploying the learned policy on physical phantoms, we test whether the representations, kinematic patterns, and temporal dependencies learned in simulation can generalize to real sensory input and real robot–tissue interaction. This stage thus represents a critical step in bridging the sim-to-real gap and evaluating whether the learned policy serves as a robust foundation for the more advanced adaptive behaviors explored in the subsequent chapters.

The contribution of this sections are:

- **Structure-Aware Trajectory Transformer.**

  We propose a hybrid neural architecture that fuses monocular RGB (optionally augmented with stereo/monocular depth estimates) with autoregressive history for the cutting tool.

The policy is built on a causal Transformer backbone that explicitly models the long-horizon dependencies of continuous cutting trajectories. A cross-attention mechanism dynamically weights visual features against kinematic history, enabling the policy to rely more heavily on proprioceptive structure when visual signals are degraded (e.g. partial occlusion). By adopting a Transformer rather than a diffusion policy, we obtain deterministic, low-latency control compatible with surgical-rate feedback loops, while still leveraging the sequence modeling capabilities validated in SRT, BC-Z, and RT-X [49–51]

- **Self-Supervised Trajectory Pretraining.**

  To address data scarcity and covariate shift, we introduce a self-supervised pretraining stage that operates purely on kinematic trajectories collected across diverse kidney–tumor scenarios (simulation and phantoms), before any visual fine-tuning. Using physics-inspired augmentations—time warping, global SE(3) transformations, and noise injection—we train the Transformer to reconstruct or predict future tool poses, thereby learning a latent manifold of feasible surgical motions. This motion-centric pretraining serves as a strong inductive bias: during IL fine-tuning with vision, the policy is biased toward smooth, anatomically plausible cuts, even when visual cues are ambiguous or partially out-of-distribution.

- **Domain-Generalization Benchmark for RAPN Cutting.** We construct and release a RAPN-inspired benchmark that explicitly varies kidney and tumor geometry, tumor location, and visual appearance across training and test sets, including simulated and physical phantoms. Unlike prior work that reports performance only on seen geometries, we evaluate our model on held-out phantoms with unseen curvatures and tumor placements, as well as on domain-shifted visual conditions. We compare against standard BC, LSTM-based policies, and vanilla Transformers without trajectory pretraining. Our experiments show that the proposed structure-aware pretraining significantly reduces trajectory error and margin violations on unseen geometries, outperforming these baselines and providing empirical evidence that geometric/kinematic pretraining is key to overcoming domain gaps in autonomous surgical cutting.

Taken together, these contributions argue that robust autonomous tumor cutting can be learned from limited expert data if the training pipeline (i) explicitly accounts for geometric and domain variability, and (ii) leverages the inherent geometric structure of surgical motion as a first-class representation. In doing so, our work closes part of the methodological gap between large-scale Transformer-based IL in general robotics and the current generation of task-specific, domain-sensitive controllers in surgical robotics, and it provides a principled path toward clinically relevant, domain-generalizable autonomy in RAPN.

2.3.2 Methods Overview

The proposed approach aims to learn visuomotor policies for robotic tumor dissection through imitation learning from demonstrations collected with surgical phantoms. Each demonstration provides synchronized RGB-D visual input and end-effector poses of the robot performing the task of tumor removal from a kidney model. The objective is to learn a multimodal policy that predicts the next end-effector pose from image observations and historical motion, enabling autonomous execution of cutting trajectories.

The method is designed to generalize across domains that share the same object categories (kidney, tumor, surgical tools) but differ in spatial configuration, geometry, color, and size. This domain adaptation capability allows the model to transfer learned behaviors across environments that preserve semantic structure while varying appearance and layout, which is an essential property of surgical environments.

Our framework consists of four main components: (1) dataset creation, processing and multi-modal data alignment, (2) a vision-trajectory transformer model that fuses visual and kinematic features through cross-attention, (3) autoregressive training with teacher-forcing and scheduled sampling, and (4) evaluation on unseen surgical scenes.

2.3.3 Dataset Creation

A total of 150 demonstrations were collected using a UR5 robotic manipulator equipped with a cauter-like surgical tool. The dataset was designed to emulate the real surgical task of tumor excision from a kidney surface, while maintaining experimental controllability and repeatability. To achieve realistic motion patterns, the robot was teleoperated during data acquisition: the first ten demonstrations were performed by experienced surgeons to capture expert motion characteristics, whereas the remaining trials were executed by trained engineers replicating those motions under supervision. This ensured consistency in tool trajectories and contact behavior while allowing a scalable collection process.

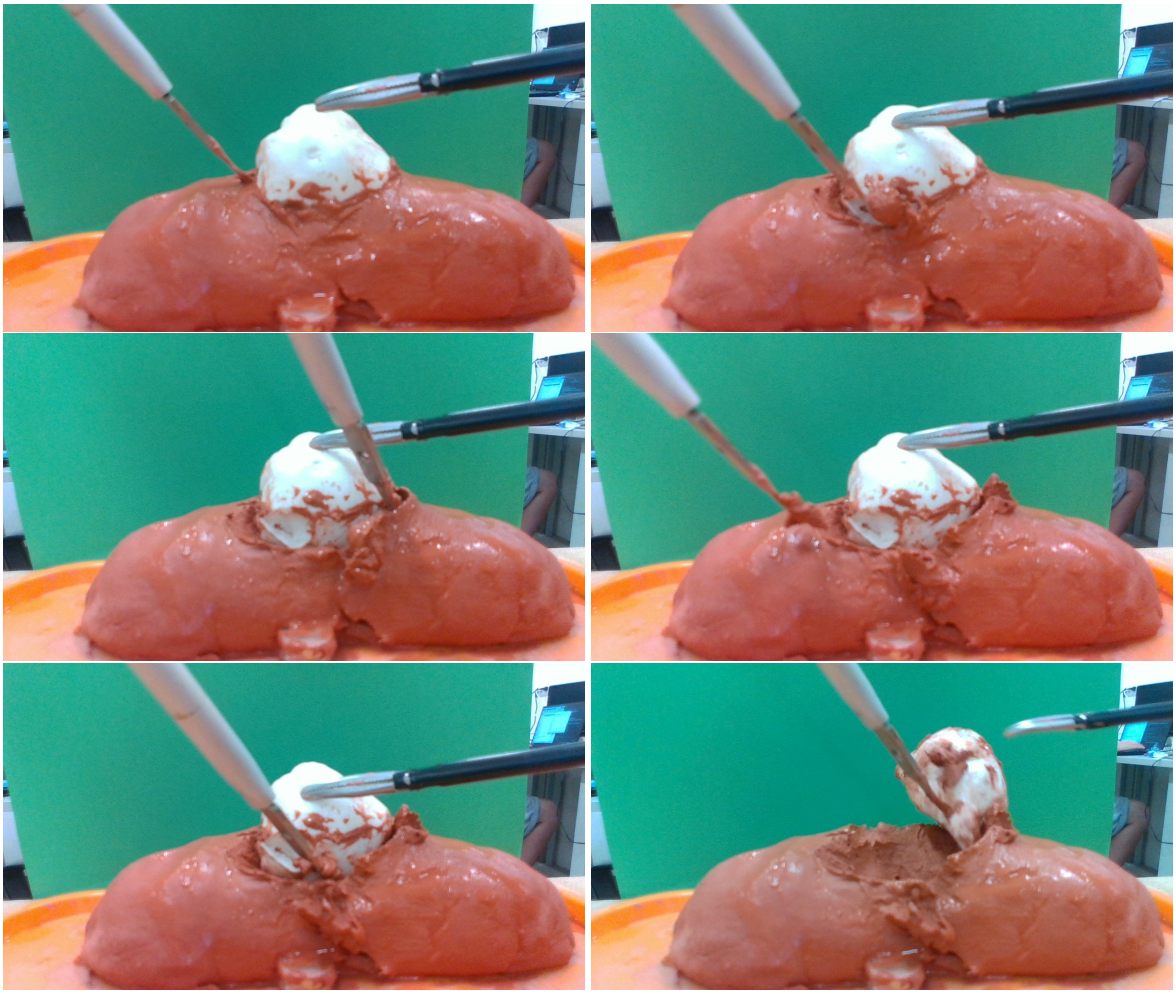


Figure 2.6: Steps of the task: the cutting tool approaches the tumor and start cutting from the left. Once it reaches the right part of the tumor it gets back to the left side and execute another cut going deeper. It repeted this process until the tumor can easily be removed.

Each demonstration records synchronized multimodal data streams: RGB images, depth maps, and full kinematic states of the robot's end effector. The RGB and depth frames were acquired from an Intel RealSense camera positioned to replicate a laparoscopic viewpoint, while Cartesian positions and quaternions were streamed directly from the robot controller. The data

are time-aligned and stored as sequences with a frame rate of 30 Hz, enabling the reconstruction of complete visual and kinematic trajectories for each demonstration.

A central goal of the dataset design is to foster domain adaptation and generalization. Although each sequence involves the same class of objects (kidney, tumor, and tools), the environments differ substantially in appearance and configuration. Across demonstrations, the kidney and tumor vary in shape, size, color, surface texture, and spatial orientation and position, while lighting and background conditions are also altered. These variations intentionally induce domain shifts that force the learning algorithm to capture task-relevant invariances, such as tool–tissue interaction dynamics and trajectory structure, rather than memorizing scene-specific features. As a result, the learned policy must adapt its motion generation to different spatial contexts while preserving the surgical intent of the action.

Despite this variability, all trajectories remain structurally consistent with the surgical cutting motion. The tool approaches the tumor, follows its boundary while maintaining appropriate contact pressure, and retreats after completion. The paths are therefore distinct yet learnable, exhibiting smooth and continuous movements representative of real surgical maneuvers. This combination of realistic teleoperated control, multimodal perception, and domain-varied repetitions provides a robust foundation for studying visuomotor imitation learning under distributional shift.

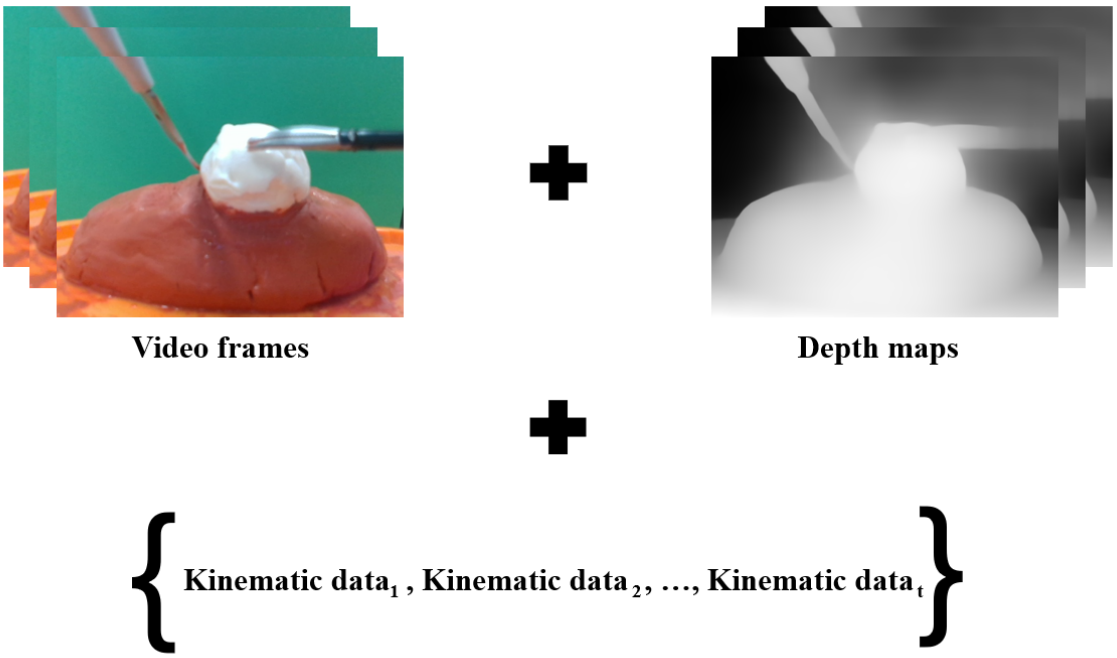


Figure 2.7: Composition of single element of dataset: RGB-D images and kidnematic data of the robot expressed in robot base frame

.

Each demonstration sequence comprises RGB frames, optional depth maps, and the corresponding robot end-effector poses recorded at 30 Hz. The raw data are organized in folders containing images and CSV files with time-synchronized Cartesian positions and orientations (quaternions). The `SequenceDataset` class preprocesses this data by (1) loading and resizing frames to 224×224 pixels, (2) normalizing intensity and depth, and (3) aligning each image with the corresponding pose entry. Missing frames or mismatched timestamps are discarded.

To improve spatial consistency, a *tool-tip correction* is applied by translating each end-effector

pose along its local z-axis to estimate the cauter tip position:

$$\mathbf{p}_{tip} = \mathbf{p}_{ee} + R(\mathbf{q}_{ee}) \cdot [0, 0, \delta_z]^T, \tag{2.6}$$

where $R(\mathbf{q}_{ee})$ is the rotation matrix derived from the end-effector quaternion and δ_z the calibrated offset of the cauter length.

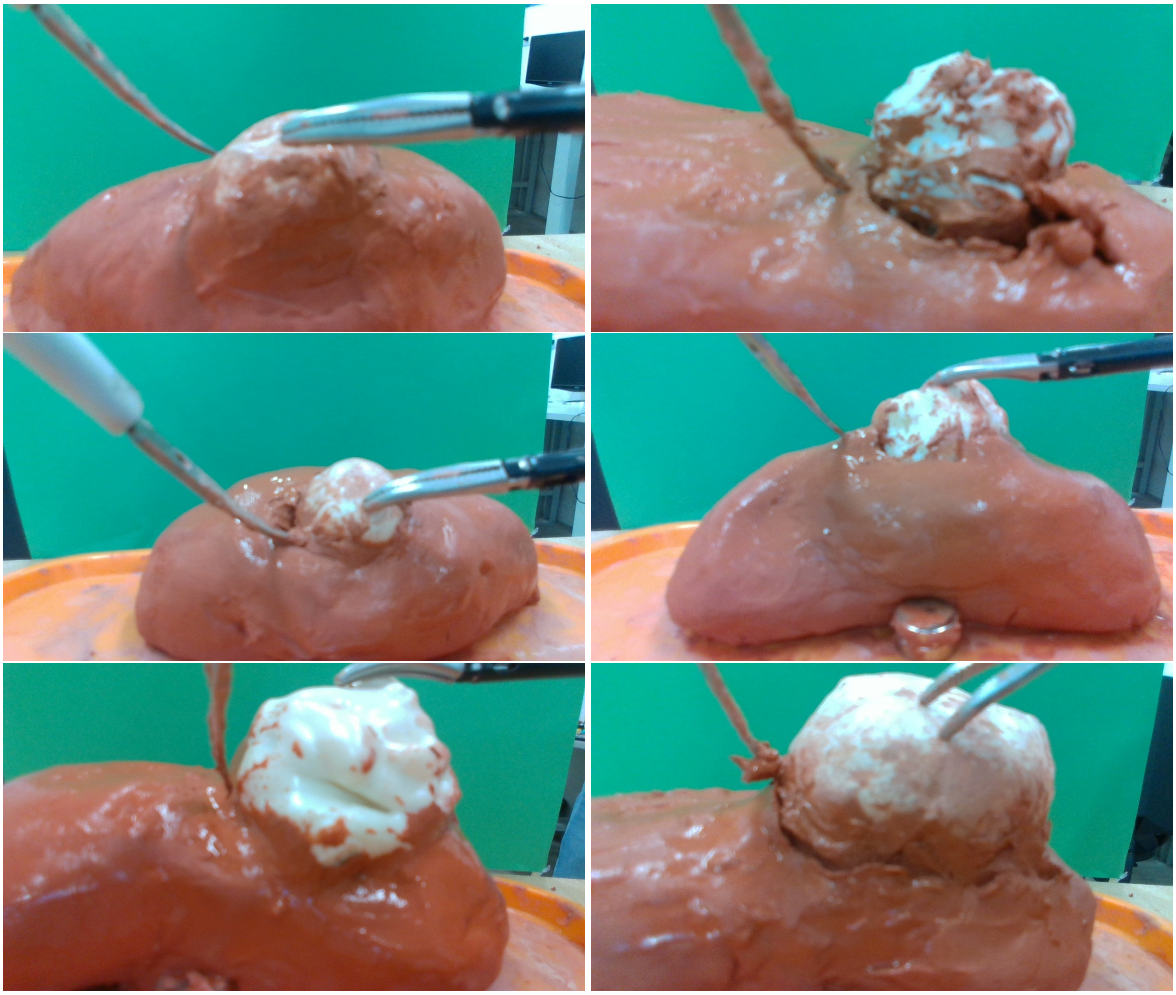


Figure 2.8: Domain shifts across different demonstration: different geometries, translation, shape, size and color of the organs (kidney and tumor); dirtiness of the tools, different perspective of the camera, different occlusion.

2.3.4 Model Architecture

The proposed architecture integrates a visual encoder, cross-modal attention, and a causal transformer for motion prediction. The architecture is depicted in Fig. **??** and described below.

to be implemented

Visual Encoder A ResNet-18 backbone extracts spatial features from each RGB (or RGB-D) frame, producing a feature map of dimension $(512, H', W')$. A linear projection maps these features into a token embedding of dimension $d_v = 256$, followed by layer normalization. Temporal consistency across consecutive frames is modeled by a lightweight 1D convolutional encoder over the last $K = 5$ frames, yielding temporally-aware image tokens.

Cross-Attention Fusion At each timestep t, the current pose embedding $\mathbf{h}_t^p$ attends to the visual tokens of the past K frames through multi-head cross-attention:

$$\mathbf{h}_t^{ctx} = \mathrm{Attn}(\mathbf{h}_t^p, \mathbf{H}_{t-K:t}^{img}), \tag{2.7}$$

where $\mathbf{h}_t^{ctx}$ represents the visual context relevant to the robot's current configuration. A gating module modulates this fusion via a learned scalar $\alpha_t \in [0, 1]$, emphasizing informative frames:

$$\mathbf{h}_t^{fuse} = [\mathbf{h}_t^p \parallel \alpha_t \mathbf{h}_t^{ctx}], \tag{2.8}$$

where $\parallel$ denotes concatenation. The fused representation is linearly projected and temporally encoded.

Causal Trajectory Transformer A causal transformer operates on the fused sequence $\{\mathbf{h}_t^{fuse}\}_{t=1}^T$, enforcing temporal causality via an attention mask that limits each token's receptive field to the previous W steps. The output embeddings are passed through a regression head to predict the next normalized pose $\hat{\mathbf{x}}_{t+1}$. An auxiliary head predicts the incremental displacement $\Delta\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_t$ directly from the visual context to improve local consistency.

2.3.5 Training of the model

Self-Supervised Model Pretraining Pretraining is particularly important in this setting because the target behavior is highly specialized: the robot must follow smooth, curved cutting paths that remain close to the tumor boundary while respecting the geometry of the kidney. Generic motion datasets (e.g., random reaching or industrial pick-and-place) are poorly matched to this structure, whereas the trajectories collected for tumor cutting are both task-specific and kinematically rich. We therefore pretrain the trajectory-transformer component of our model in a self-supervised manner directly on these surgical-like trajectories, using only robot poses and no visual input. This provides a strong prior over feasible surgical motions and reduces the amount of supervised visuomotor data required for stable long-horizon prediction.

Let $\mathbf{x}_t \in \mathbb{R}^7$ denote the end-effector pose at time t, parameterized as Cartesian position and unit quaternion, $\mathbf{x}_t = [p_t^x, p_t^y, p_t^z, q_t^x, q_t^y, q_t^z, q_t^w]^\top$. Given a trajectory $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, the self-supervised objective trains the causal transformer to predict the next pose from its history:

$$\hat{\mathbf{x}}_{t+1} = f_\theta(\mathbf{x}_{1:t}), \quad t = 1, \ldots, T-1, \tag{2.9}$$

where f_θ shares the same temporal architecture as in the full multimodal model but operates only on pose tokens. The loss is defined as a Huber regression over the full pose vector,

$$\mathcal{L}_{\mathrm{SSL}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathcal{L}_{\mathrm{huber}}\big(\hat{\mathbf{x}}_{t+1}, \mathbf{x}_{t+1}\big), \tag{2.10}$$

with quaternions renormalized after each update to remain on the unit hypersphere. This training phase is fully self-supervised, requiring only kinematic logs of the teleoperated demonstrations. After convergence, the weights of the trajectory transformer are used to initialize the corresponding module in the multimodal model and subsequently fine-tuned jointly with the visual encoder and cross-attention layers.

To increase robustness and encourage invariance to small variations in execution, we apply a set of SE(3)-aware data augmentations directly to the pose sequences during pretraining. First, we inject local stochastic perturbations by adding zero-mean Gaussian noise to Cartesian positions and by slightly jittering tool orientations in Euler angle space before converting back to normalized quaternions. Second, we apply global rigid transformations, including rotations about the main axes and mirroring operations, which rotate or reflect the entire trajectory

around an origin while consistently updating orientations. These operations preserve the relative shape and continuity of the cutting path but expose the model to different global poses and viewpoints.

Temporal structure is further diversified through time-domain augmentations. We perform smooth time stretching and compression of trajectories by resampling them to longer or shorter lengths using linear interpolation for positions and spherical linear interpolation (SLERP) for quaternions. Additional augmentations introduce mild velocity jitter, piecewise non-uniform time warps, and waypoint dropout with interpolation, which randomly removes intermediate waypoints and reconstructs them from neighboring poses. Finally, spatial scaling and low-pass filtering with high-frequency noise are used to simulate subtle variations in curvature and path smoothness while maintaining physical plausibility.

All augmentations are designed to preserve the high-level semantics of the task: the tool approaches the kidney surface, follows a closed or semi-closed path around the tumor, and then retracts. By training the transformer to predict the next pose across this enriched family of trajectories, the model learns a domain-specific prior over feasible surgical motions that is invariant to small spatial, rotational, and temporal perturbations.

Training Objective The network is trained using a combination of absolute, velocity, and consistency losses:

$$\mathcal{L} = \mathcal{L}_{abs} + \alpha\mathcal{L}_{\Delta} + \beta\mathcal{L}_{cons} + \lambda_{aux}\mathcal{L}_{aux}, \tag{2.11}$$

where $\mathcal{L}_{abs}$ penalizes the distance between predicted and ground-truth poses, $\mathcal{L}_{\Delta}$ enforces accurate motion increments, and $\mathcal{L}_{cons}$ encourages temporal consistency between predicted and integrated motions. Each term employs the Huber loss:

$$\mathcal{L}_{huber}(x, y) = \begin{cases} 0.5(x-y)^2, & |x-y| < \delta, \\ \delta(|x-y| - 0.5\delta), & \text{otherwise.} \end{cases} \tag{2.12}$$

Scheduled sampling progressively replaces ground-truth poses with model predictions according to a probability schedule p_{TF} decaying over epochs, bridging the gap between teacher-forcing and autoregressive inference.

2.4 Results and Experiments on Physical Phantoms

The performance of the imitation learning framework was evaluated on a set of twelve previously unseen phantom environments, each presenting variations in tumor geometry, phantom appearance, lighting conditions, or relative organ-tool configurations. This evaluation aimed to assess the ability of the learned policy to generalize beyond the demonstrations and to execute the surgical subtask with sufficient accuracy for subsequent autonomous control.

2.4.1 Quantitative Results

Across all successful trials, the policy achieved an average positional error of approximately 0.2 cm with respect to the expert demonstrations. This level of accuracy is consistent and stable across phantoms of different shapes and colors, confirming that the model is capable of extracting task-relevant geometric and kinematic patterns even under moderate domain shifts.

Orientation prediction proved more challenging, with an average angular deviation of 10° between predicted and ground-truth tool orientations. While this error remains acceptable for the class of tasks considered in this stage, particularly given the limited orientation variability in the demonstrations, it indicates that orientation learning requires further refinement as task complexity increases.

Overall, the policy successfully executed the task in 9 out of the 12 test environments, corresponding to a 75% generalization success rate. Failures occurred primarily in phantoms

with extreme geometric variations or highly unusual tumor placements. In these scenarios, the model struggled to maintain accurate orientation and occasionally deviated from the expected spatial trajectory.

2.4.2 Qualitative Observations

Qualitatively, the learned trajectories exhibit smooth and continuous motion, closely resembling the expert demonstrations. The model reliably reproduces the multi-stage structure of the surgical task, including approach, contact, alignment, and cutting phases. In several cases, the predicted motion was smoother than the demonstrations themselves, consistent with the temporal smoothing effects introduced during learning.

Notably, even in challenging scenarios with visual artifacts or illumination changes, the model was able to maintain consistent predictions, reflecting robust feature extraction from RGB-D data. However, strong changes in phantom coloration or large deviations in tumor geometry occasionally reduced prediction stability, particularly in orientation.

2.4.3 Future Experimental Steps

The next stage of this research involves deploying the learned policy on the physical robotic system. These real robotic experiments will evaluate the model under realistic execution constraints and dynamic interaction with tissue-mimicking phantoms. Upcoming work will integrate the actuated surgical tool, enabling the robot not only to follow trajectories but also to grasp and manipulate silicone organs.

Future experiments will focus on:

- validating the predicted trajectories during real tool–tissue interaction,
- assessing how positional and orientation errors propagate during closed-loop execution,
- evaluating robustness under physical perturbations, tool deflection, and occlusion,
- testing execution across a larger set of phantoms and more complex surgical subtasks.

These developments will provide the necessary validation of the imitation learning framework in realistic conditions and pave the way for its integration with the deformation-aware reasoning and perception modules presented in the following chapters.

2.4.4 Considerations and Future Steps

In this stage of the research, the primary goal was to transfer an imitation learning policy, trained from expert demonstrations, onto a real robotic system operating on physical, tissue-mimicking phantoms. This represents a significant milestone in the advancement of autonomous surgical robotics, marking one of the first points in the thesis where learned surgical skill is executed outside simulation and within a real, deformable environment. The successful execution of the learned trajectories on silicone phantoms demonstrates a tangible step toward overcoming the Sim-to-Real gap, validating that data-driven policies can extend beyond virtual training and perform meaningfully in contact-rich physical scenarios.

The central achievement of this research stage is the demonstration that an imitation learning policy, originally trained using expert demonstrations, can accurately reproduce surgical-like trajectories on a physical robot without requiring additional tuning or calibration. The policy achieved a high degree of precision, with an average positional error of approximately 0.2 cm during task execution. This level of accuracy is remarkable considering the differences in visual conditions, geometry, and noise properties between training and testing domains. It provides strong evidence that the network internalized the underlying structure of the trajectory generation task, rather than memorizing training patterns tied to a specific environment.

Another notable outcome is the demonstration of domain adaptation. The model was tested on phantoms differing slightly in appearance, material coloration, tumor geometry, or surface reflectivity, and it maintained robust performance throughout. This suggests that the learned representations are not overly sensitive to superficial visual variations and that the policy can tolerate moderate domain shift, a requirement for any future application in clinical environments, where patient anatomy and visual conditions naturally vary.

While the results of this research stage are promising, several limitations must be acknowledged. First, despite the use of physical phantoms, the experimental setup does not fully capture the anatomical and biomechanical complexity of real soft tissue. The phantoms, although geometrically derived from CT scans, remain homogeneous and isotropic compared to actual organs. They lack vessel networks, heterogeneous stiffness distributions, tissue anisotropy, blood perfusion, and dynamic physiological motion such as breathing or pulsation. Consequently, while the demonstrated generalization is meaningful, it remains constrained to a simplified tissue environment.

A second limitation concerns the end-effector used during this stage of experimentation. The tool mounted on the robot is not yet a fully actuated surgical instrument and lacks the ability to grasp, retract, or manipulate tissue. It cannot open or close, does not transmit significant forces, and cannot perform layered or incremental cuts. As such, the demonstrations and policy execution are limited to pure end-effector trajectory reproduction, without any dynamic tool-tissue interaction such as traction or spreading. This constraints the complexity of tasks that can be autonomously executed, and limits the ecological validity of the experiments.

The cutting task itself, while visually complex, is simplified in terms of material dynamics, tool-tissue friction, and environmental constraints. Silicone phantoms offer a controllable surface but do not impose the same mechanical resistance or nonlinear response characteristic of biological tissue. Real tissue often requires variable pressure, dynamic angle adjustment, and adaptive selection of cutting depth; none of these were required in the current experiments, meaning the system has not yet been exposed to conditions requiring active adaptation of cutting mechanics.

In addition to these structural limitations, there are algorithmic ones. Although the model demonstrated strong positional accuracy, it is clear from the results that further refinement is necessary. The predicted trajectories could be smoother, better reflecting the micro-adjustments seen in expert motions. More importantly, orientation predictions (roll, pitch, and yaw) exhibited higher variance and occasional instability, which could compromise the fidelity of fine surgical actions. These issues are likely due to the limited size and diversity of the demonstration dataset, particularly regarding orientation changes during cutting.

Despite these limitations, the path forward is clearly defined and highly actionable. Improving performance and generalization begins with expanding the dataset. This involves collecting more demonstrations across a wider range of phantoms, lighting conditions, viewing angles, and trajectory types. By broadening the representational space, the model becomes more robust to variations and reduces the potential for covariate shift during execution. Additional emphasis should be placed on capturing more complex tool motions, including non-linear orientation sequences and trajectories involving multi-phase actions.

Next, advancing the neural network architecture is a crucial direction. Modern visuomotor policy architectures such as transformer-based models, cross-modal fusion networks, or diffusion-policy-based autoregressive predictors have shown extraordinary potential in robotic manipulation tasks. These architectures may offer better temporal coherence, improved handling of long-horizon sequences, and better ability to capture subtle spatial relationships within surgical scenes. Integrating such models could significantly improve both accuracy and smoothness of the generated trajectories.

From the hardware perspective, future experiments should transition toward more realistic phantoms that mimic heterogeneous tissue properties and offer more complex interactions. Using phantoms with embedded vessels, layers of tissue, or variable stiffness regions would help bridge

the gap between simplified tests and more realistic surgical conditions. Ultimately, the findings from this stage will be strengthened by integrating the actuated surgical tool developed in later stages of the thesis. This will allow the system to perform more complex actions such as grasping, retraction, suction, or clamping, rather than merely following a predefined cutting path. The ability to coordinate multiple tools or multi-phase actions will be essential as the research expands toward procedures requiring bilateral coordination, such as selective clamping or layered dissection.

Finally, the scope of tasks must be expanded. Future work should aim to move beyond simple, single-phase cutting, toward full multi-tool workflows. Examples include retraction, counter-traction, estimation of tissue tension, sequential tumor resection, and autonomous vascular clamping. Increasing the repertoire of autonomously executable tasks will test the versatility of the policy and push the learning framework toward more general, adaptive surgical autonomy.

In summary, this research stage marks an essential step toward the full realization of autonomous surgical systems for soft tissue manipulation. The successful transfer of an imitation learning policy from virtual training to real-world robotic execution validates the foundation of the learning component in the broader autonomy framework. While limitations exist, they provide a natural and productive roadmap for the next phases of research. The future steps identified, improving data richness, upgrading neural architectures, developing realistic phantoms, integrating advanced tools, and expanding task complexity, lay the groundwork for increasingly capable, safe, and intelligent autonomous surgical systems.

2.5 Conclusion: Learning from Demonstrations, Imitating and Taking Decisions Through Experience

The work presented in this chapter established the learning foundation of the thesis by developing and validating an imitation learning framework capable of acquiring expert-level surgical trajectories and transferring them to both virtual and physical robotic systems. The research followed a two-phase progression. First, policies were developed and refined within a virtual environment, where large-scale data, controlled experimental conditions and rapid iteration enabled systematic exploration of architectural choices. Second, these policies were deployed on real robotic phantoms, demonstrating successful execution and generalization across changes in visual appearance and geometry. Together, these two phases form a cohesive investigation into how robots can learn surgical skill directly from human demonstrations, providing the essential competence upon which all subsequent reasoning and perception modules rely.

Within the broader Learning, Reasoning, Perceiving and Acting (LRPA) framework that structures the thesis, the role of this chapter is centered on learning. The work accomplished here yields the policy function , which encodes the expert strategy for performing the tumor-cutting task and, in later extensions, related subtasks. This learned policy represents the distilled blueprint of expert behavior, describing how the robot should move, orient its tool, and progress along the desired trajectory. Without such a learned model, the subsequent modules—deformation-aware perception, geometric reasoning for clamping or cutting, and physical execution through the actuated tool—would have no foundation on which to operate. This chapter therefore contributes the essential skill substrate that enables the robot not only to imitate but also, later, to adapt. It provides the initial competence that subsequent perception and reasoning modules will refine and correct based on the real-time state of the surgical environment.

A structured comparison of the two research phases highlights the strengths and limitations of this approach.

The virtual environment phase offered several important advantages. First, it enabled large-scale data generation at negligible cost. Thousands of trajectories could be recorded automatically, providing diverse state–action pairs essential for training deep neural networks. Second,

the controlled nature of simulation allowed rigorous testing and debugging without the practical constraints of hardware, sensor noise or experimental failures. Third, simulation provided perfect ground truth for tool position and orientation, making it possible to evaluate and optimize the model's positional accuracy during early development. These strengths made simulation an ideal starting point for refining the architecture, selecting the temporal window size and evaluating different training strategies.

However, the virtual phase also exhibited clear limitations. The most significant was the sim-to-real gap: the mismatch between simulated and real environments that prevents policies trained solely in simulation from transferring to physical systems without degradation. Visual conditions in simulation are highly idealized, lacking the noise, reflections, shadows and complex textures present in real scenes. Additionally, the physics models used to simulate tool–tissue interactions are simplified approximations of real soft-tissue mechanics. They cannot capture anisotropic stiffness or friction properties that strongly influence real surgical execution. Consequently, although simulation was invaluable for early-stage learning and model selection, it could not serve as the sole source of training data.

The second phase, focused on real robotic phantom validation, addressed many of these limitations. The most important achievement was demonstrating the successful execution of the learned policy on a physical robot interacting with silicone phantoms. The system achieved an average positional error of approximately 0.2 cm, a strong benchmark for this class of surgical tasks. Moreover, the policy demonstrated promising domain adaptation: it generalized across phantoms with different shapes, tumor boundaries or surface textures, maintaining stable execution even under moderate distribution shift. These results provide concrete evidence that the model captured underlying geometric and kinematic patterns that extend beyond individual training samples.

The tool used in this phase was not an actuated surgical instrument and lacked the ability to open, close or manipulate tissue. As a result, the executed trajectories were purely positional and did not include active cutting dynamics, tissue retraction or variable force profiles. Furthermore, the phantoms, although derived from patient CT data, were mechanically homogeneous and lacked layers, vessels or biological variability. Consequently, although the imitation learning policy succeeded in following the path, it did so under simplified mechanical conditions that do not fully capture the complexity of real surgical tasks.

The network's ability to reproduce orientation accurately was also limited. While positional tracking was strong, the orientation prediction occasionally exhibited jitter or drift, suggesting that further architectural refinement is needed to handle rotational components of the motion. Additionally, the relatively small size of the physical demonstration dataset, compared to the virtual dataset, constrained generalization and smoothness.

These limitations define a clear and actionable roadmap for the future development of the learning pillar.

Future improvements should focus on enhancing the performance and generalization of the learned policies. Incorporating more advanced neural architectures, such as transformer variants specialized for long-horizon trajectories, diffusion policies for smooth motion generation, or hybrid vision–language models with multimodal conditioning, may significantly improve robustness to distribution shift. A specific focus should be placed on orientation learning, potentially through auxiliary orientation losses, quaternion regularization strategies, or the inclusion of orientation-based attention modules.

A key direction for bridging the reality gap will be the integration of explicit domain adaptation strategies. Approaches such as domain randomization, adversarial feature alignment or cross-modal consistency constraints could reduce visual and geometric discrepancies between simulation and reality. Simultaneously, the virtual environment should be enhanced with more realistic rendering, physics-based simulation of deformable organs and adverse visual conditions to reduce the reliance on physical data.

Future work must also expand the scope of tasks beyond single-path cutting. Once integrated with the actuated surgical tool, the learning framework should be extended to grasping, clamping and multi-tool coordination. Learning-based recovery strategies, in which the robot detects deviations from the desired trajectory and autonomously corrects them, are essential for surgical safety. Techniques such as hierarchical imitation learning, conditional policy switching or diffusion-based corrective control may contribute to the development of robust failure recovery.

Finally, the data collection process should be enhanced. A larger and more diverse dataset, including different phantoms, lighting conditions, trajectory families and tool–tissue interactions, will strengthen policy generalization. Capturing demonstrations at varying skill levels or with deliberate perturbations could also help train models that are resilient to unexpected states. An important next step for advancing the learning pillar is the transition from training exclusively on controlled simulation and phantom demonstrations to incorporating real surgical video data. The foundation developed so far, simulation for large-scale supervised learning and physical phantoms for Sim-to-Real grounding, provides a strong starting point for this transition. The learned models already possess a structured understanding of the surgical scene, the temporal dynamics of the task and the mapping between visual appearance and tool motion. These capabilities significantly reduce the difficulty of learning from real surgical videos, which are inherently more variable, noisier and lack perfect action labeling.

Future research should focus on leveraging the existing learned representations as priors when training on real clinical footage. This can be achieved through semi-supervised or weakly supervised learning strategies in which the model first learns visual embeddings and temporal patterns from unlabeled surgical videos and then aligns these representations with the demonstration-based control policy. Domain adaptation methods, including contrastive learning across phantom and real frames, synthetic-to-real style transfer and temporal alignment techniques, will be critical for bridging visual disparities between laboratory setups and the operating room.

Incorporating real surgical videos will not only increase the ecological validity of the learned policies but also enable the model to internalize a richer set of behaviors, anatomical contexts, tissue appearances and unpredictable events. Ultimately, this step will move the learning component closer to the clinical reality, positioning the system for future integration into more advanced reasoning, perception and control modules capable of operating safely in true surgical environments.

In summary, this chapter establishes the learning foundation of the entire thesis. Through virtual and physical experimentation, it demonstrates that imitation learning can effectively capture and reproduce expert surgical trajectories, serving as the initial competence upon which adaptive perception and reasoning modules are built. While limitations remain, they provide a clear roadmap for enhancing the learning component and integrating it into increasingly complex, realistic and safe autonomous surgical systems.

While this chapter has demonstrated that robots can learn expert-derived trajectories and reproduce them with high accuracy in both virtual and physical settings, learning alone is not always sufficient for autonomy in soft-tissue surgery.

This limitation motivates the need for the next stages of the thesis, which introduce the remaining pillars of the LRPA framework: reasoning, perceiving and acting. Reasoning is essential for transforming the static policy learned in this chapter into a dynamic, context-aware behavior that can update target points and trajectories based on the evolving state of the surgical scene. Perception is required to estimate this evolving state in the first place; without real-time tracking of soft-tissue deformation, no adaptive correction can occur. Acting, which becomes increasingly important as manipulation complexity grows, requires precise and reliable hardware capable of executing context-aware actions with fine control over tool motion and force application.

The transition from this chapter to the next therefore reflects a natural progression: having

endowed the system with a foundational surgical skill through imitation learning, the subsequent chapters focus on equipping the robot with the capabilities required to use that skill safely and intelligently in real surgical environments. The integration of learning with perception, reasoning and advanced actuation forms the core of the thesis and ultimately defines the pathway toward adaptive and autonomous soft-tissue surgical robotics.

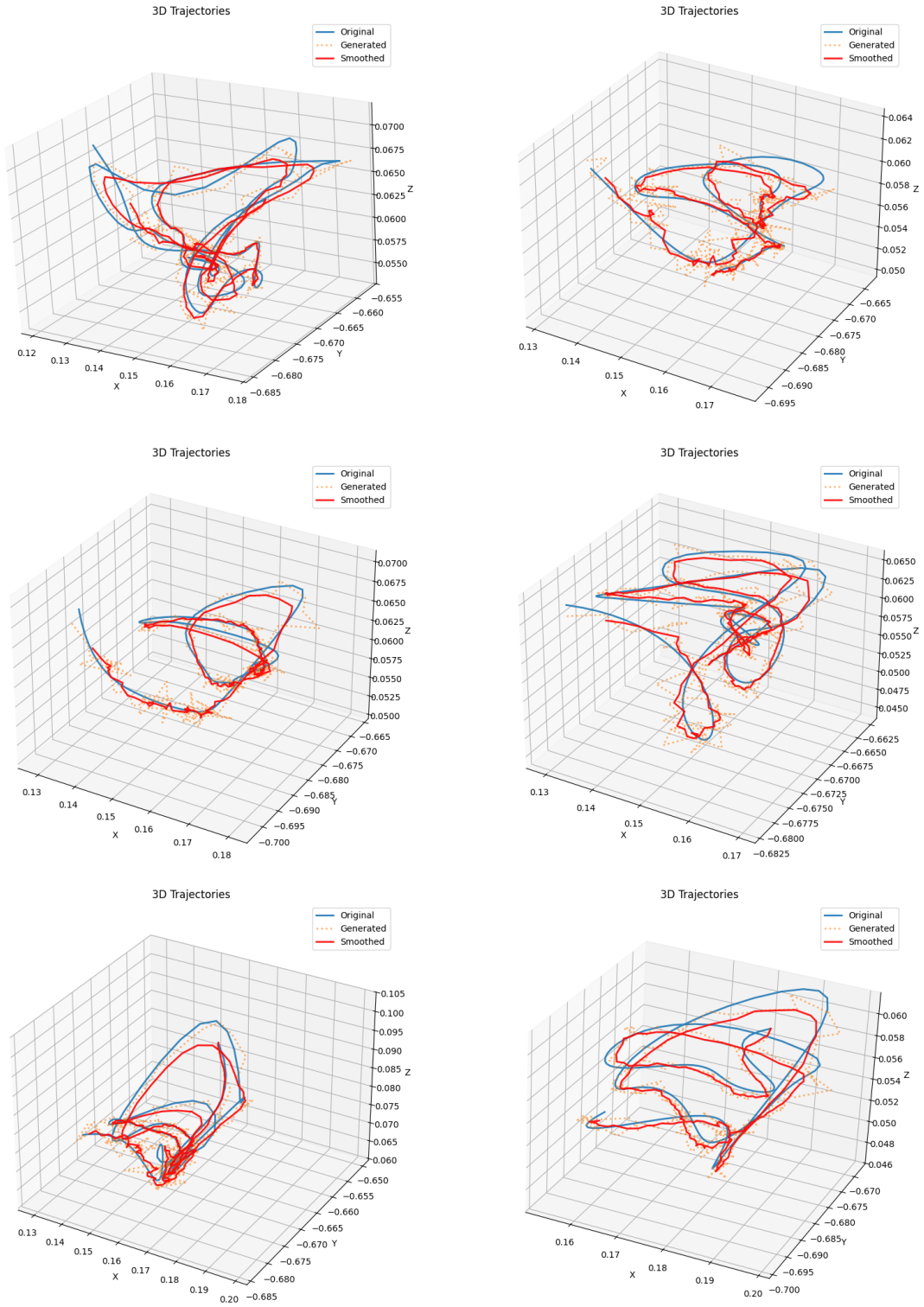


Figure 2.9: Plot comparison between GT trajectories (blue), generated trajectories (organge) and filtered generated trajectories (red)

Chapter 3

From Learned Decisions to Surgical Reasoning: Planning for Selective Clamping

3.1 Introduction: Surgical Reasoning

While learning from demonstrations equips the robot with motor competence and implicit decision patterns, genuine autonomy in surgery demands the ability to reason. Surgical reasoning involves understanding the anatomical context, evaluating multiple possible strategies, predicting the consequences of different choices, and selecting the course of action that optimally aligns with clinical objectives such as safety, preservation of healthy tissue, and procedural efficiency.

In a real surgical environment, reasoning is the mechanism by which a surgeon determines what to do, why to do it, and in which order. These decisions integrate anatomical knowledge, task goals, surgical constraints, pathology-specific considerations, and an understanding of risk. Reproducing this cognitive layer in robotic systems is essential if we aim to move beyond motion replication toward systems capable of meaningful and independent surgical assistance. Imitation learning provides an essential foundation, but it cannot fully capture the deliberative processes that guide a surgeon's actions when faced with anatomical variability, task complexity, or unexpected intraoperative conditions.

Robotic systems must therefore be capable of formulating task-level decisions rather than simply executing motions. They must identify relevant structures, evaluate alternative strategies, and select actions that optimize a clinical objective. This requirement becomes particularly evident in procedures that rely on complex planning and anatomical reasoning rather than pure execution. Among such procedures, selective clamping during robot-assisted partial nephrectomy represents a compelling and clinically relevant example. Selective clamping requires the identification of arterial branches supplying blood to a renal tumor, followed by a deliberate choice of clamping points that minimize ischemic damage to healthy parenchyma while ensuring adequate devascularization for safe tumor resection. Unlike trajectory following, this task is inherently decision-driven: the surgeon must choose *where* to clamp, not simply *how* to perform the mechanical action.

This chapter focuses on the computational and algorithmic foundations of surgical reasoning in the context of selective clamping. The goal is not merely to automate the mechanical act of clamping but to reason about vascular anatomy and determine which arteries should be occluded to achieve an optimal surgical outcome. By modeling the arterial network as a graph, quantifying blood flow distribution, and formulating ischemia minimization as an optimization problem, the system can autonomously derive a clamping plan that reflects a deliberate surgical decision, rather than a predefined rule or demonstration.

More broadly, this represents a shift from low-level action prediction to high-level planning.

While the robot in the previous chapter learned how to cut a tumor based on human demonstrations, the reasoning framework introduced here allows the robot to determine what surgical action must be taken in the first place. This distinction marks a crucial evolution in autonomy: learned skills provide a vocabulary of actions, but reasoning determines how those actions should be arranged to construct a coherent and safe surgical procedure.

Surgical reasoning also addresses a fundamental limitation of purely learned systems: the inability to generalize to novel anatomical configurations. Whereas imitation learning thrives on repeated patterns, surgical reasoning thrives on structural understanding. By analyzing the specific anatomy of each patient, the robot can identify key structures, compute relationships between them, and derive a sequence of actions tailored to that particular case. Such patient-specific planning is a core requirement in surgery, where anatomical variations are the norm rather than the exception.

Selective clamping serves in this chapter as a representative example because it encapsulates the essential elements of surgical reasoning: anatomical interpretation, goal optimization, constraint satisfaction, and planning under uncertainty. However, the implications of this work extend far beyond clamping alone. Any surgical task that requires evaluating alternatives, anticipating outcomes, or selecting strategic points of intervention can benefit from similar reasoning mechanisms. Surgical suturing, vessel isolation, tumor margin definition, dissection planning, and access optimization are all tasks in which decision-making plays a critical role and where robotic autonomy could substantially enhance clinical outcomes.

The integration of surgical reasoning into robotic systems is therefore indispensable for achieving the next level of autonomy. It requires mathematical modeling, algorithmic formulation, and an explicit representation of surgical objectives. More importantly, it bridges the gap between learning and perception. Reasoning provides structure but also highlights its own limitations: it assumes that anatomy is static, that pre-operative models remain valid throughout the procedure, and that planned decisions remain optimal as surgery unfolds. These limitations reveal the need for the next component of autonomy: perception-driven adaptation.

Thus, this chapter establishes surgical reasoning as a necessary layer between learned skills and perceptual awareness. It shows how a robot can transition from reproducing expert actions to generating patient-specific plans that optimize surgical objectives. Selective clamping is introduced not as an isolated problem but as a case study that crystallizes the role of decision-making in robotic surgery. The detailed algorithmic framework is presented in the following sections, setting the stage for the integration of reasoning with real-time deformation awareness in the subsequent chapter.

3.2 Problem Definition: Selective Clamping in Partial Nephrectomy

Partial nephrectomy is considered to be the gold standard treatment for localized renal tumors [52] and the use of robot-assisted systems has been increasingly adopted for executing the procedure (Robot-Assisted Partial Nephrectomy or RAPN) [53]. Typically, the procedure is performed with interruption of renal blood flow with clamping of the renal arteries, especially for large and anatomically complex masses with deep parenchymal invasion. The execution of the vascular clamping allows the surgeon to work in a relatively bloodless field, facilitating tumor resection and closure of the parenchymal defect; however, the temporary interruption of arterial flow may lead to ischemic injury of the renal parenchyma. As a consequence, the interruption of the flow and the related ischemia should be as short as possible and should involve the smallest renal area. Indeed, the preserved healthy renal parenchyma nearby the tumor is one of the most important predictors of long-term renal function, whereas strong evidence concerning the ischemic damage to renal function are still lacking [54]. In order to maximize the functional advantage of partial nephrectomy, different ischemia approaches have been proposed. In partic-

ular cold or warm ischemia, selective and super-selective ischemia and no-ischemia (or clampless) approaches were published [55]. Among all these different techniques, selective clamping allows to induce ischemia only in the immediate tumor feeding artery, minimizing the ischemia of the healthy parenchymal tissue that surrounds the tumor. However, performing selective clamping on the basis of Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) images only could be very difficult, since the surgeon could be unable to identify the exact intrarenal vascular anatomy and to predict the real tumor blood supply from the images [56]. In order to provide assistance to the surgeon in identifying the tumor blood supply, 3D virtual models of the patient anatomy have been proposed to be examined pre-operatively [57]. An additional technological step has been the superimposition of the 3D virtual model directly on the real-time images showed in the robotic console during RAPN [58], [54]. Thanks to the superimposition, the surgeon can directly see the 3D virtual model projected on the patient anatomy intra-operatively. However, in these previous works, the choice of the arterial branches to be clamped is still left to the surgeon, that examines the anatomy of the patient to identify the best clamping solution. Moreover, to the best of author's knowledge, researches on the development of automatic solutions to plan the clamping procedure and integration with the robotic systems used for RAPN are still lacking. The development of an algorithm for automatically computing the optimal clamping points would be a further step towards a precise and completely assisted surgery and it could be exploited as a planner for the robotic system. In general, the robotic system could be either a robotic assistant for the surgeon, as proposed in European projects like SARAS [59], [60] or one arm of the robotic system that is used for executing the RAPN, suitably adapted.

In this section, we propose the first step towards the above mentioned automatic and integrated solution. In particular, we present an algorithm to automatically compute the clamping points, based on the description of the arterial tree as a graph and on the resolution of an optimization problem. The final goal is to provide to the surgeon the clamping points which allow to minimize the ischemia percentage related to the patient's specific anatomy, thus preserving the health of the remaining renal parenchyma. To the best of author's knowledge, this is the first paper where the computation of the clamping points is computed automatically. Moreover, we show how to exploit the proposed algorithm as a planner for a robotic system to perform a clamping procedure. The output of our algorithm produces the clamping points the robot should reach for implementing the clamping procedure. For each clamping point, a desired pose for the robot is computed in such a way that the clamping tool is orthogonal to the vase to be clamped and that collision with other organ is avoided. The proposed architecture is finally validated on different patient anatomies and using a robotic setup.

3.3 Proposed System Architecture

The proposed system architecture, summarized in Fig. 3.1, consists of an algorithm for automatically computing the clamping points and a robotic system for executing the selective clamping in practice. In particular, the algorithm for *clamping point computation*, exploits the 3D model reconstructed from a Computed Tomography with Angiography (CTA) and extracts the tree of the arterial branches and an associated graph to compute the clamping points that allow to stop the arterial flow flowing towards the tumor while minimizing the percentage of ischemia. Then, these points are used by the robotic system for the automatic *robotic execution* of the selective clamping task.

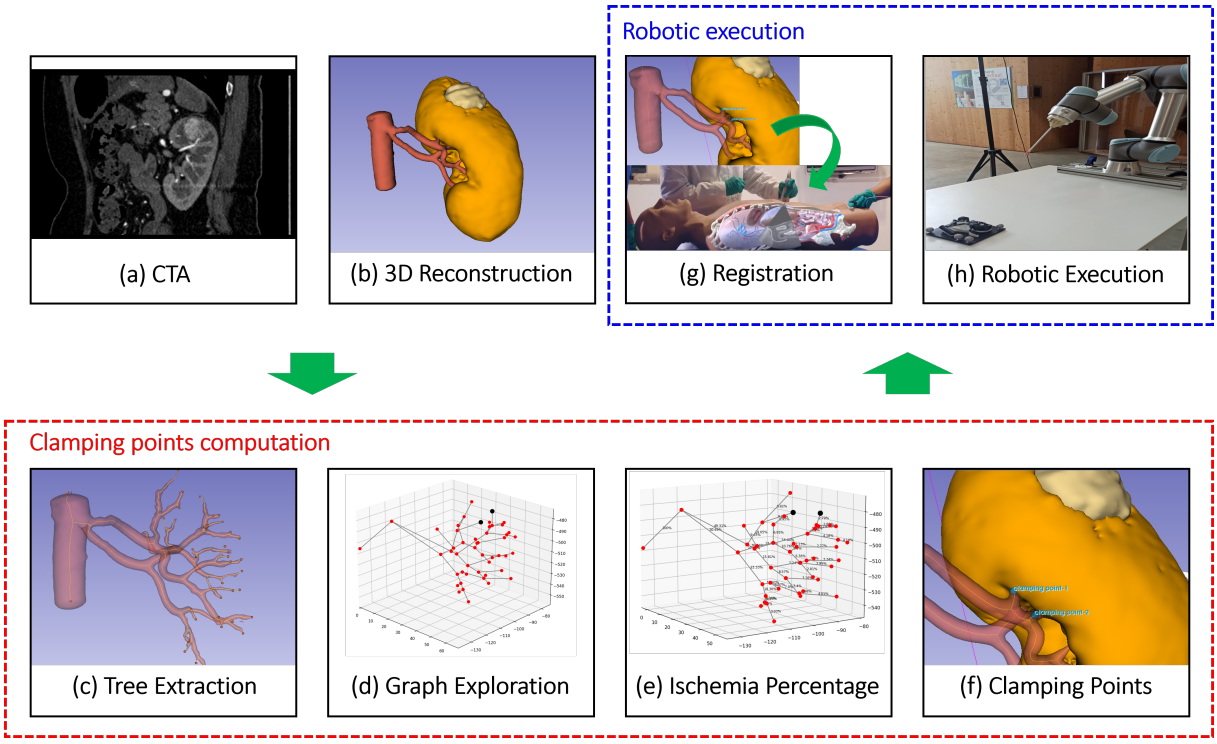


Figure 3.1: Overall architecture.

3.3.1 Clamping points computation

The clamping point computation algorithm relies on a 3D reconstruction of anatomical structures from pre-operative CTA images. From the CTA (Fig. 3.1(a)), a 3D model of the kidney, tumor, and renal arteries branching off from the aorta is generated (Fig. 3.1(b)). Using Hounsfield Units for segmentation, the 3D reconstruction follows the procedure in [61]. The vascular segmentation used to initialize the selective clamping pipeline was obtained with TotalSegmentator, an nnU-Net-based whole-body CT segmentation framework. Since manual ground-truth vessel annotations were not available for our acquired cases, we report the quantitative performance provided by the original TotalSegmentator validation as a reference baseline. In their Radiology: Artificial Intelligence study, the authors report a mean Dice similarity coefficient of 0.943 on a held-out test set over 104 anatomical structures, including eight vascular classes, and a significant improvement over a publicly available baseline model in a direct comparison (Dice 0.932 vs 0.871). These values provide an estimate of the expected segmentation quality when using TotalSegmentator as an automated initialization.

Semi-automatic artery reconstruction. While TotalSegmentator provides a robust segmentation for major vessels, the reconstruction of the full renal arterial tree is not fully automatic in our workflow. In particular, smaller distal branches and low-contrast portions of the tree may require manual refinement (e.g., local editing of masks and/or manual completion of missing segments) before centerline extraction. This semi-automatic step was introduced to ensure a topologically consistent arterial model suitable for downstream reasoning (branch identification and clamping-point localization). A dedicated quantitative evaluation of the vessel masks and extracted centerlines on our data will require expert-validated ground truth annotations and is left as future work.

With the reconstructed 3D model, the algorithm automatically determines clamping points to halt arterial flow toward the tumor while preserving kidney functionality. Arterial branches are mathematically described through tree extraction and graph creation (Fig. 3.1(c)). Graph

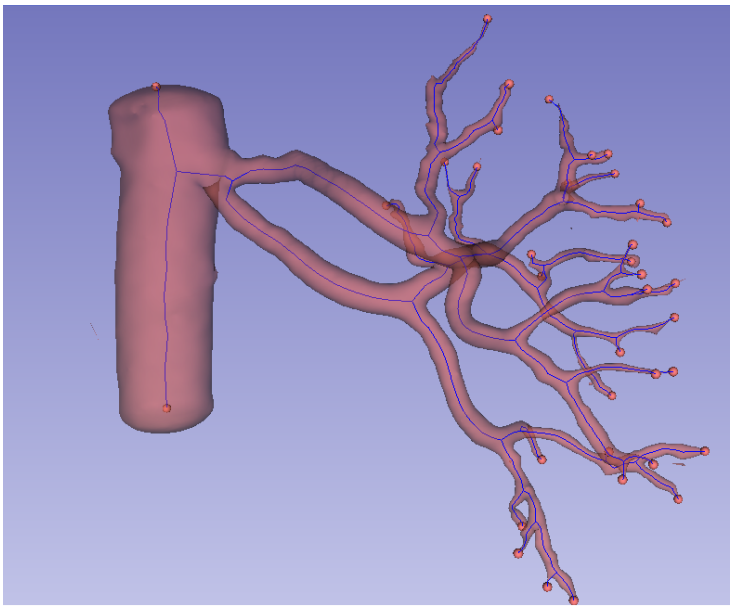
search and optimization techniques are then applied to compute clamping points (Fig. 3.1(d)-(f)). Subsequent sections detail each step of the algorithm.

Tree Extraction:

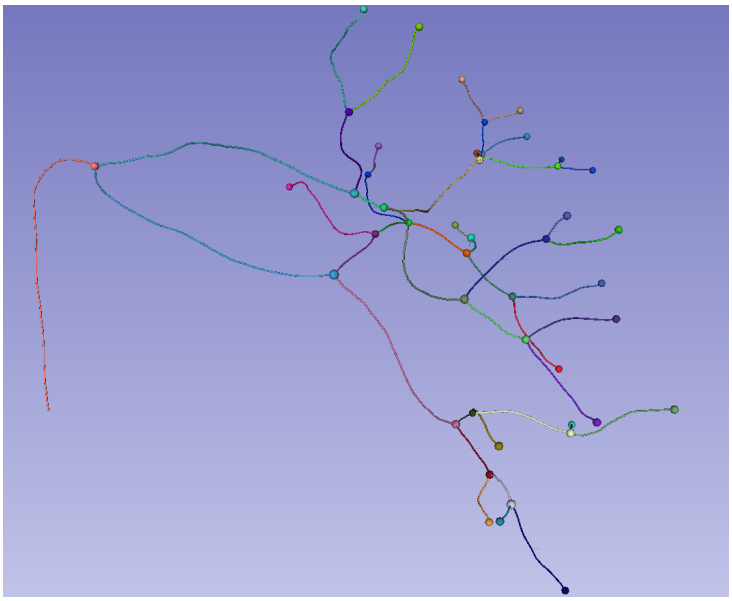
Starting from the 3D model of the arteries, the algorithm computes the centerline, defined as the line maximizing the distance from the boundary between two sections of the tubular structure (Fig. 3.2(a)). Various techniques exist for centerline computation in blood vessel geometry characterization [62, 63]. This paper adopts the approach in [64], where centerlines are paths on Voronoi diagram sheets minimizing the integral of the radius of maximal inscribed spheres. This yields information about the blood vessel diameter, crucial for computing ischemia percentage (Section 3.3.1).

To compute clamping points, a mathematical description of arterial branches is needed. The arterial structure is modeled as a tree graph.

Let us define $\mathcal{G}^k$ as the tree graph which will model the arterial tree involved in the bedewing of the kidney, where $\mathcal{V}\left(\mathcal{G}^k\right)$ and $\mathcal{E}\left(\mathcal{G}^k\right) \subset \mathcal{V}\left(\mathcal{G}^k\right) \times \mathcal{V}\left(\mathcal{G}^k\right)$ are the vertex set and the edge set, respectively. The edges represent the arteries while the vertices represent the points where the arteries divide. Thus, we computed the branches of the tree (edges) and the vascular bifurcations (vertices), by following the strategy suggested in [65]. Each centerline point corresponds to the radius of the maximum inscribed sphere, enabling the construction of a tube around the centerline with the envelope of these spheres. Tubes intersections reveal vascular bifurcations, establishing branches and nodes for the arterial tree [65] (Fig. 3.3). Each color represents a distinct arterial branch, and nodes are formed at branch intersections.



((a)) Extracted arterial tree.



((b)) Graph representation of arterial branches.

Figure 3.2: Example of arterial tree extraction and graph creation for clamping computation.

To focus on arterial branches flowing into the tumor, the algorithm identifies paths from the aorta to the tumor by computing intersections between the arterial tree and the tumor (yellow and green regions, Fig. 3.3). Branches corresponding to these intersections are determined by their minimum distance from the centroid of each intersecting region. These identified branches' leaves are considered of interest, as they are endpoints within the tumor, marking the terminus of the arterial tree.

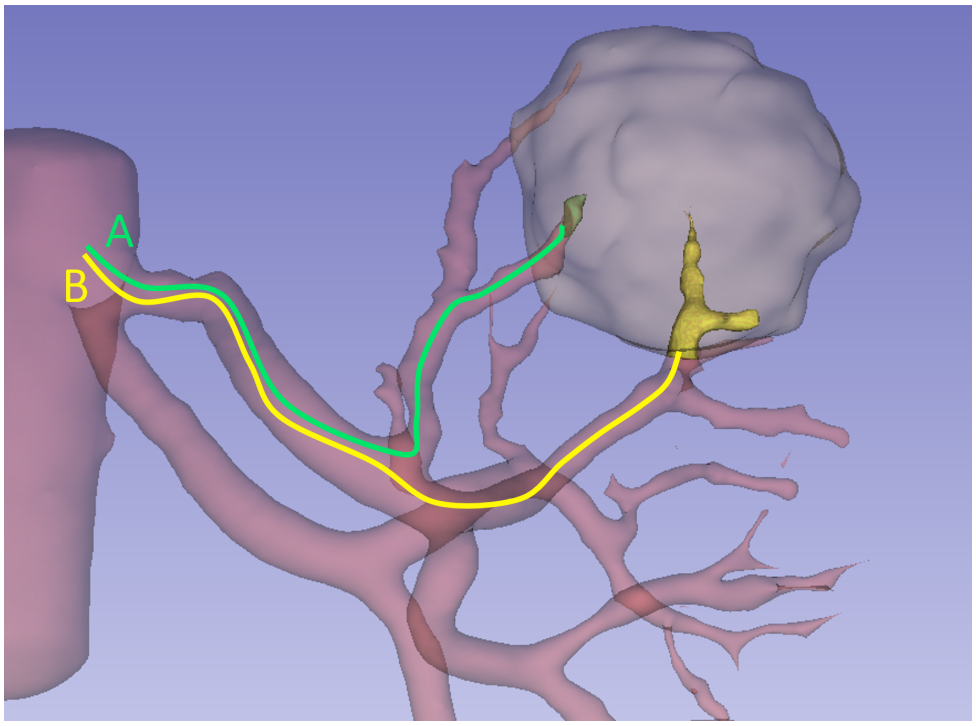


Figure 3.3: Branches of the arterial tree which end up into the tumor and regions of intersection between the arterial tree and the tumor.

Graph Exploration:

Once having identified the leaves of interest, the goal is to find the paths that start at the initial node, corresponding to the aorta, and end into the tumor, which means the leaves of interest themselves (green path A and yellow path B in Fig. 3.3). These paths will correspond to the only arterial branches which provide blood to the tumor and that will be interested by the clamping procedure.

In order to formally identify the specified path, we derived the tree graph $\mathcal{G}^k$ by applying the graph exploration method named as Breadth-first [66]. Figure 3.4 shows the graph associated to the tree in Fig. 3.3. In the figure, each vertex represents a bifurcation point of the arteries, while each edge represents the branch which is generated starting from the bifurcation point. The black vertices represent the leaves that end into the tumor, as computed before.

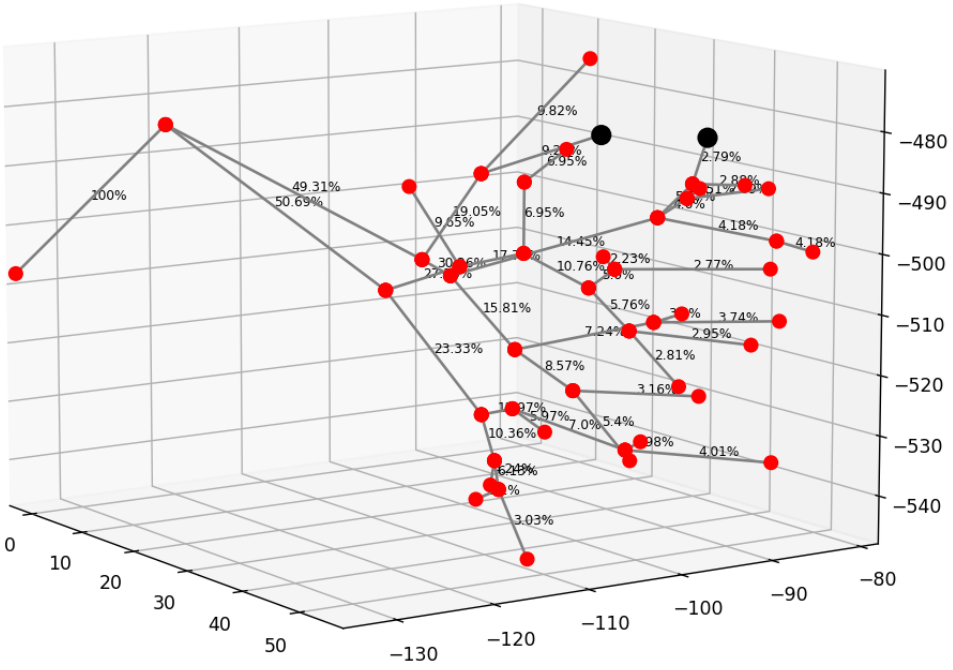


Figure 3.4: 3D graph which is computed starting from the arterial tree and ischemia percentage associated to each arc.

Ischemia percentage computation:

Once having created the graph $\mathcal{G}^k$ starting from the arterial tree, the following step is to compute the ischemia percentage related to each edge of the graph, which means the ischemia which would be induced in the kidney if the clamping is executed on that arteria. The computation of the ischemia percentage must take into consideration the diameter of each arteria. Then, we will formulate an optimization problem to compute the clamping points, according to our model.

Let $A \in \mathbb{R}^{n \times n}$ be the adjacency matrix of the graph $\mathcal{G}^k$: each element a_{ij} is 1 if an edge exists from node i to node j, zero otherwise. Then, let $F_j \in \mathbb{R}$ be the percentage of the blood that flows into a generic node j, with $j \in 1, \dots, n$. Since all the blood that flows into the tree comes from the aorta artery, considering it as node 1 means that:

$$F_1 = 1.0 \tag{3.1}$$

while the other flows can be computed as:

$$F_j = \sum_{i=1}^{n} a_{ij} F_i \frac{d_{ij}}{\sum_{k=1}^{n} a_{ik} d_{ik}} \tag{3.2}$$

where d_{ij} is the diameter of the edge from i to j if exist, or zero otherwise. It is worth noting that, since $\mathcal{G}^k$ is a tree graph, given j, only one of the a_{ij} for $i \in 1, \dots, n$ is different from zero, individuating the parent node of node j. This means that $\sum_{i=1}^{n} a_{ij} F_i$ defines the flow coming from the parent node, while (3.2) represents the same flow weighted as a function of the diameter of the arteries starting from the parent node.

The ischemia percentage p_{ij}^k associated to the edge (i, j), that represents the percentage of the total blood flowing into the node j, can be computed as:

$$p_{ij}^k = a_{ij}^k F_j^k \tag{3.3}$$

where F_j^k is the flow in the kidney graph $\mathcal{G}^k$ and a_{ij}^k is 1 if an edge from node i to node j exists in the kidney graph, zero otherwise. Figure 3.4 shows the existing branches of the tree graph $\mathcal{G}^k$ with the ischemia percentage associated to each edge.

Clamping Points Computation:

The final goal of the algorithm is to provide to the surgeon the clamping points which allow to minimize the total ischemia percentage related to the specific anatomy of the patient, thus preserving the healthy of the remaining renal parenchyma. The tree graph $\mathcal{G}^k$ represents the arterial tree of the whole kidney, that means both the branches of arteries which are outside the kidney and the branches inside the kidney. However, the clamping procedure can be performed only on branches which are outside the kidney. To this aim, we computed a new 3D model, which is the difference between the overall reconstruction and the reconstruction which involves the parenchyma and the tumor, and we implemented the optimization problem for computing the clamping points on the new generated model, which contains only the branches of arteries outside the kidney. Then, let us define $\mathcal{G}^t$ as the subgraph of $\mathcal{G}^k$ involved in the bedewing of the tumor such that the edge (i, j) can be clamped, i.e. is located outside the kidney. The interested edges are part of the aorta-tumor paths computed from the tree ending in the black leaves depicted in Fig. 3.4. Let $X^k \in \mathbb{R}^{n\times n}$ be the adjacency matrix for which each element x_{ij}^k is 1 if the edge between the node i and j is clamped or 0 otherwise. The total ischemia percentage p^k can be then computed as:

$$p^k = \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^k p_{ij}^k \tag{3.4}$$

This computes the amount of kidney tissue that will not receive blood during a surgical procedure where certain branches are clamped. The percentages of ischemia for each clamped branch, i.e. the percentage of blood that is provided by each clamped branch, are summed to obtain the total percentage of kidney tissue that will be affected by the lack of blood flow.
Furthermore, when performing a clamping operation on the kidney, it is important to ensure that the procedure is executed outside of the kidney, to avoid any physical damage or interference with the organ. As a result, it becomes necessary to locate the edges, or part of them if they are partially inside the organ, that are outside of the kidney. This allows to find the edges where the operation can be physically executed. The external edges can be found by computing the difference between the 3D model of the arteries and the 3D model of the kidney.
The solution for the selective clamping procedure consists in finding the edges, among the edges outside the kidney and the ones that feeds the tumor, to clamp, such that the ischemia percentage is minimized and the flow of blood to the tumor is interrupted. This can be done by solving the following optimization problem:

$$\begin{aligned} &\min_{X^k} \; p^k(X^k) \\ &s.t. \; \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^t a_{ij}^t F_j^t = 1.0 \end{aligned} \tag{3.5}$$

where a_{ij}^t is 1 if the edge from node i to j exists in the tumor graph, 0 otherwise; x_{ij}^t is 1 if the edge between the node i and j in the tumor graph is clamped, 0 otherwise; F_j^t is the flow in the tumor graph. The constraint ensures that the total blood inflow to each node must equal the total outflow from that node. In other words, the total flow from the aorta is equal to the total blood flow into the kidney. Additionally, the constraint imposes that at least one branch must be clamped, otherwise the operation is not executed. By identifying the values that satisfy the constraints and minimize the objective function, the problem provides a quantitative approach to determine the course of action that interrupts the blood flow to the tumor while minimizing ischemia and maintaining blood flow to healthy regions of the kidney.

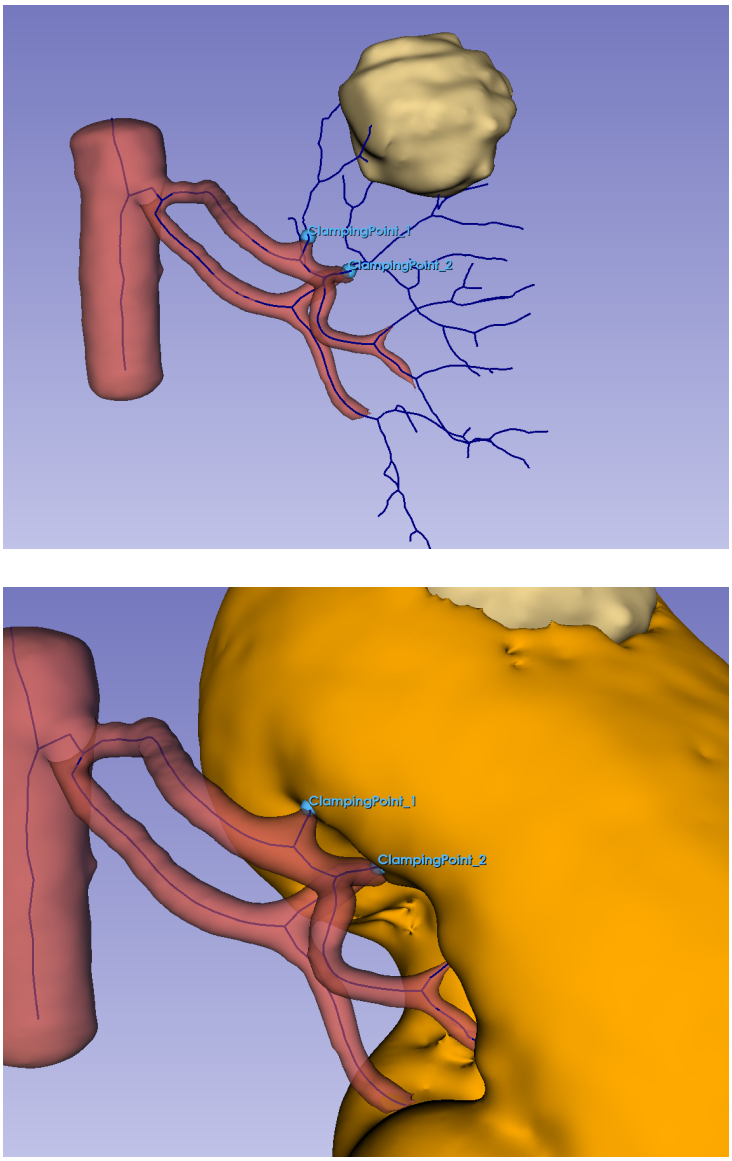


Figure 3.5: clamping points computed according to the optimization problem.

3.3.2 Robotic Execution of Selective Clamping

Once the clamping points have been computed as a result of the algorithm proposed in Section 3.3.1, the execution of the selective clamping procedure could be performed automatically by a robotic system. To this aim, two additional steps are required. In particular, at first the 3D model reconstructed on the basis of pre-operative CTA must be aligned on the real body anatomy. This is actually a non-trivial problem, since organs deform and move during the surgery, due to patient positioning, organ dissection or movement, etc. However, since the focus of this section is not tissue rigid and non-rigid deformation, which we will disccuss about later on, we will consider a rigid transformation between pre-operative and intra-operative phase for the moment. It is important to note that the scenario being considered in our work is intentionally simplified and not entirely realistic (Fig. 3.1(g)). Then, the clamping poses are computed so that the clamping tool is orthogonal to the vase and the robot avoids collision with other anatomical structures. Then, the robotic system can execute the procedure, moving to the points computed by the algorithm and performing the clamping 3.1(h)).

Patient-3D model registration

The three-dimensional reconstruction of the anatomical structures has its own reference system, which coincides with that of the CTA machine since the model is segmented starting from the CTA images. Consequently, the clamping points are also returned with respect to the same reference system. Therefore, the registration between the reference system of the model and that of the robot base is necessary to allow the robotic execution of the clamping procedure.

This is a known problem in the literature and several methodologies have been developed to solve it (see for example [67]). Since this is not the focus of this work, we assume to have the poses of the model and of the base of the robot available in real-time, provided by, for example, an external tracking system.

Let ${}^{w}T_m \in \mathbb{R}^{4\times4}$ and ${}^{w}T_r \in \mathbb{R}^{4\times4}$ be the poses of the model and of the robot base with respect to the world reference system respectively, and:

$$ {}^{m}pc_i = \begin{bmatrix} pc_i^x & pc_i^y & pc_i^z & 1 \end{bmatrix}^T \quad i \in 1, \dots, p $$

the i-th clamping point expressed in the homogeneous coordinates, with $p \in \mathbb{N}$ the number of clamping points. The i-th clamping point can be transformed with respect to the robot base reference frame as:

$$ {}^{r}pc_i = {}^{r}T_m \, {}^{m}pc_i \tag{3.6} $$

where

$$ {}^{r}T_m = ({}^{w}T_r)^{-1} \, {}^{w}T_m \tag{3.7} $$

provides the registration matrix.

Robotic execution of selective clamping

The overall execution of the selective clamping is performed providing the robot a sequence of poses such that the procedure is completed correctly.

First, a clamping pose ${}^{r}Tc_i$ needs to be computed from the clamping points ${}^{r}pc_i$ in order to drive the robot to a consistent reference.The position part ${}^{r}pc_i$ of ${}^{r}Tc_i$ coincides with the clamping point while the orientation part ${}^{r}Oc_i$ is computed by forcing the surgical tool to lie perpendicularly to the clamping artery. The orthogonality constraint would provide the most surface area contact between the vase and the jaws, which distributes pressure evenly and provide a secure clamp. We compute the plane of the vase as the plane that passes through the clamping point and whose normal is the vector that goes from the clamping point to the previous point on the branch curve. Considering the normal of the plane as the x-axis, we can assume the tool is orthogonal to the artery when its x-axis coincides with the x-axis of the artery. The orthogonality can be achieved by setting ${}^{r}Oc_i = R_z(\theta c_i)R_y(\alpha c_i)R_x(\psi c_i)$, with $\theta c_i = atan2({}^{y}nc_i, {}^{x}nc_i)$ and $\alpha c_i = -asin({}^{z}nc_i)$ where $R_z(\theta c_i)$, $R_y(\alpha c_i)$ and $R_x(\psi)$ are the rotation matrices, respectively, about the z-axis by an angle of θc_i, the y-axis by an angle of αc_i and the x-axis by an angle of ψc_i; $nc_i = [{}^{x}nc_i, {}^{y}nc_i, {}^{z}nc_i] \in \mathbb{R}^3$ is the unit vector individuating the normal of the i-th clamping point plane. The value of ψc_i is free, and represents the rotation the tool can have around the artery without affecting the orthogonality. The degree of freedom is used to fix the pose of the tool so that there are no collisions with the surrounding organs. In order to avoid collisions, we consider the clamping tool as a sequence of $N_t \in \mathbb{N}$ virtual spheres placed along a straight line and we ensure a safe distance $d_{sft} \in \mathbb{R}^+$ of each sphere from each organ. The problem is formulated as follows:

$$\min_{\psi c_i, \epsilon_t, \epsilon_a} \quad \theta = \arccos\left(\frac{\hat{z}c_i \cdot \hat{z}c_i^{des}}{|\hat{z}c_i||\hat{z}c_i^{des}|}\right)$$

$$\begin{aligned} s.t.\ & {}^{r}Oc_i = R_z(\theta c_i + \epsilon_t)R_y(\alpha c_i + \epsilon_a)R_x(\psi c_i) \\ & d_{o_i}^{tool_j} > d_{sft} \quad \forall i \in N_o \, \forall j \in N_t \\ & -\pi \leq \psi c_i \leq \pi \\ & -\epsilon_t^{max} \leq \epsilon_t \leq \epsilon_t^{max} \\ & -\epsilon_a^{max} \leq \epsilon_a \leq \epsilon_a^{max} \end{aligned} \tag{3.8}$$

where $\hat{z}c_i \in \mathbb{R}^3$ is equal to the third column of ${}^{r}Oc_i$ which is the unit vector representing the z-axis of the tool, i.e. the unit vector indicating the tool main axis, and $\hat{z}c_i^{des} \in \mathbb{R}^3$ its desired value.

Here, the term $d_{o_i}^{tool_j} \in \mathbb{R}$ indicates the distance between the i-th organ and the j-th virtual sphere, with $N_o \in \mathbb{N}$ the number of organs in the operating area, and $N_t \in \mathbb{N}$ the number of sphere built along the tool. However, it is not always possible to find a feasible solution in all scenarios and a limitation in the usability of the clamping tool may occur, since the space the robot can move in is very limited. For this reason, to make the optimisation problem less stringent, the orthogonality constraint is relaxed by the introduction of $\epsilon_t, \epsilon_a \in \mathbb{R}$, which introduce an angular deviation from orthogonality. In fact, the correct execution of the procedure can be obtained even without forcing the perfect orthogonality between the clamping tool and the considered artery. The term $\epsilon_t^{max}, \epsilon_a^{max} \in \mathbb{R}^+$ are used to limit this deviation. If no feasible solution that assures the deviation is smaller than ϵ_t^{max} and ϵ_a^{max} is found, the surgeon is warned and the operation is not executed autonomously.
Allowing some deviation from perpendicularity is not the ideal solution, but a certain degree of tolerance is accepted maintaining a reasonable level of stability and precision. This guarantees the robot to move without touching any other organs and prevent undesirable damages.

Since the robot cannot move directly between the clamping poses without injuring the patient, for each clamping pose a pre-clamping pose is computed as:

$${}^{r}\bar{T}c_i = {}^{r}Tc_i \, T_{sft} \tag{3.9}$$

where $T_{sft} \in \mathbb{R}^{4\times4}$ represents a transformation with respect to the clamping pose such that the generated pre-clamping pose ${}^{r}\bar{T}c_i$ is safe. In our specific scenario we consider to have a rod-like laparoscopic instrument with reference frame located to the tool-tip with outgoing z-axis, and we set T_{sft} as a single negative translation $z_{sft} \in \mathbb{R}^-$ along the z axis, i.e.:

$$T_{sft} = \begin{bmatrix} & & & 0 \\ & I_{3\times3} & & 0 \\ & & & z_{sft} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.10}$$

Once the pre-clamping poses are computed, the procedure is performed for each clamping point as summarized in Algorithm 1. In particular, at first the i-th clamping point ${}^{m}pc_i$ is registered in the robot base frame ${}^{w}T_r$ exploiting the registration matrix ${}^{r}T_m$ according to (3.6) (line 1). Then, in lines 1-1 the i-th clamping and pre-clamping poses are computed starting from (3.6) and according to (3.9). Then, the robot moves to the pre-clamping and clamping poses (lines 1-1), executes the clamping procedure (line 1) and returns to the pre-clamping pose (line 1) for starting the following execution.

Algorithm 1 RoboticExecution

```
0: procedure RoboticExecution(^m pc_i, ^r T_m, T_sft)
0:   for i ← 1 to p do
0:     ^r pc_i ← registerPoint(^m pc_i, ^r T_m)
0:     ^r Tc_i ← computeClampingPose(^r pc_i)
0:     ^r T̄c_i ← ComputePreClampingPose(T_sft)
0:     Robot ← moveTo(^r T̄c_i)
0:     Robot ← moveTo(^r Tc_i)
0:     Robot ← executeClamping()
0:     Robot ← moveTo(^r T̄c_i)
0:   end for
0: end procedure=0
```

Robotic Motion Planning and Execution

For each clamping point, the robot computes a target pose ensuring that the instrument axis is orthogonal to the artery to maximize contact stability. The orientation of the tool is defined by three Euler angles θ_{ci}, α_{ci}, and ψ_{ci} as:

$$R_{ci} = R_z(\theta_{ci})R_y(\alpha_{ci})R_x(\psi_{ci}) \tag{3.11}$$

where the x-axis of the tool is aligned with the normal vector of the arterial plane at the clamping location. The value of ψ_{ci} represents a free rotation about the artery axis, adjusted to prevent collisions with surrounding structures. Collision avoidance is implemented by modeling the surgical instrument as a sequence of N_t virtual spheres along its axis, ensuring a safe minimum distance d_{sft} from neighboring organs.

To ensure safety during motion, a pre-clamping pose is defined by translating the clamping pose by a distance z_{sft} along the tool's z-axis:

$${}^r\bar{T}_{ci} = {}^rT_{ci}\, T_{sft} \tag{3.12}$$

where T_{sft} is a translation matrix describing the safety offset.

The robotic controller executes the following sequence for each clamping point: move to the pre-clamping pose, approach the clamping pose, perform clamping, and return to the pre-clamping pose. This sequence guarantees both accuracy and safety.

3.4 Experimental Validation

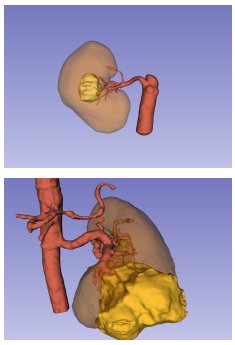
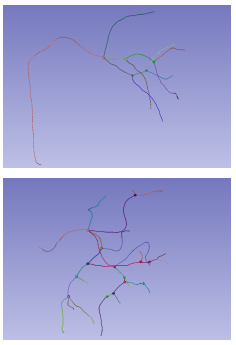
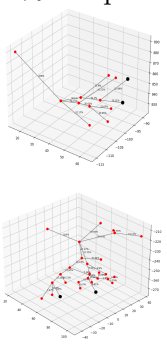
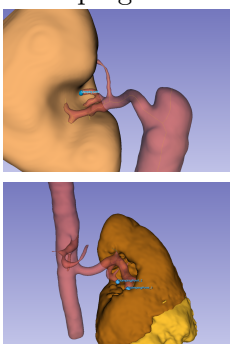
3.4.1 Validation with different CTA

In order to validate the system, at first we applied the automatic procedure described in Section 3.3.1 on the anatomy of additional patients with respect to the one taken as example throughout the paper. In particular, Table 3.1 shows the implementation of the different steps of the algorithm for computing the clamping points executed on two of the total twelve different anatomies the model has been tested on. We acquired the CTA of the patients and segmented it to generate the 3D reconstruction of the anatomy. Then, the 3D reconstruction is provided as input to the algorithm, which automatically generates the arterial tree and the tree graph, computes the ischemia percentage associated to each edge and solve the optimization problem to compute the clamping points to be provided to the robotic system.

3.5 Robotic execution of selective clamping

Once having validated the algorithm for clamping points computation on different patients, we implemented and emulated the automatic execution of the selective clamping procedure on a

Table 3.1: Validation of the algorithm for clamping points computation on different patients.

Segmented CTA	3D Reconstruction	Arterial Tree	% Graph	Clamping Points
				

robotic setup. As a use case, we considered the anatomy of the patient shown in Fig. 3.1 and reconstructed starting from the CTA (Fig. 3.1(a)). At first, we printed a 3D support containing the model of the kidney with the tumor. We used an Universal Robot UR5e endowed with a 3D printed emulated laparoscopic scissor instrument, and the Optitrack as external tracking system. In order to properly track the kidney model and the robot base, both were equipped with a set of 3D markers. The experimental setup can be clearly seen in the accompanying video clip.

The robotic controller is implemented as a Finite State Machine (FSM), in order to correctly execute and synchronize all the steps of the procedure, and to make it flexible with respect to the clamping points number. Figure 3.6 reports the results of the experimental validation. In particular, Fig. 3.6(a) shows the positions reached by the robot and the pre-clamping and clamping points, while Fig. 3.6(b) shows the triggering of the FSM. As can be seen in Fig. 3.6(b), the system starts in the WAIT_FOR_START state where the robot is located in the middle of the operational area waiting for a command. Once the clamping points are computed and the surgeon trigger the clamping execution the state transits to MOVE_TO_PRE_CL ($t = 5s$). Then, the robot starts moving towards the first pre-clamping pose ${}^r\bar{T}c_0$, as can be seen in Fig. 3.6(a). When the robot reaches the desired pose ($t = 8s$), the system switches to the MOVE_TO_CL state, and the robot moves to the first clamping pose ${}^rTc_0$. Then, the controller state transits to CL and the clamping is executed ($9s \leq t \leq 10$). Since we do not have a real kidney and a real clamping instrument, the clamping phase is simulated only as a waiting time, emulating the time required for a real clamping phase. Finally the state is set as RETURN_TO_PRE_CL where the robot come back to the pre-clamping pose ${}^r\bar{T}c_1$ in order to guarantee safety.

Operations from MOVE_TO_PRE_CL to RETURN_TO_PRE_CL are repeated for all the clamping points, which in the specific case of this validation are 2. When all the clamping points are clamped ($t = 16s$), the state transits back to WAIT_FOR_START.

3.6 Discussion and Future Directions

Partial nephrectomy is typically performed with interruption of renal blood flow with clamping of the renal arteries. However, the interruption of arterial flow generates ischemia of the involved renal parenchyma. To avoid excessive renal damage, selective clamping can be applied in order to minimize the part of the renal parenchyma interested by ischemia. In this paper we proposed an algorithm for automatic computing clamping points that allow to minimize the ischemia percentage. Then, a robotic system is used for emulating and automatically executing the clamping procedure. However, autonomous surgical procedures are a complex and challenging area that requires further research to overcome limitations and achieve practical solutions. Our current work has some deviations from real-world scenarios, including the absence of accounting

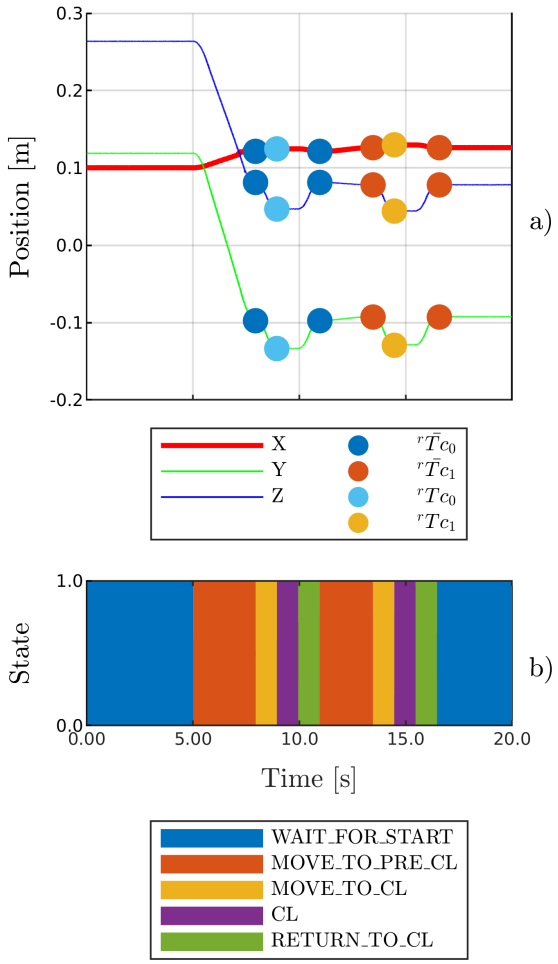


Figure 3.6: Experimental results. (a) Cartesian position of the tool-tip of the clamping instrument, with the pre-clamping and the clamping poses. (b) State of the robot controller during the clamping experiment.

for organ deformation and the use of rigid transformations between pre-operative and intra-operative phases. Later on we will focus on addressing these limitations by integrating real-time registration of 3D models with the patient's body to superimpose the 3D model on the real anatomy. This will allow visualization of clamping points directly on the surgeon's console during the procedure in real-time. Additionally, future works will consider the possibility for the robot to move anatomical parts to reach surgical goals, and an automated way to compute 3D models from CTAs will be used. Furthermore, in-silico organs and fluids will be used in future experiments, and the computation of clamping tool pose and trajectory will be optimized to ensure optimal solutions. Currently, only kidney, artery, pancreas, and spleen are considered as obstacles to be avoided, and future research will aim to expand this scope. Moreover, we aim at commanding one arm of the robotic system used for executing RAPN, in order to perform the selective clamping procedure automatically.

3.7 Conclusions: Toward Contextual Autonomy

The approach presented in this chapter establishes a concrete step toward autonomous decision-making in robot-assisted partial nephrectomy, focusing on the problem of selective clamping. By combining anatomical reconstruction, graph-based modeling of the arterial tree, and an

optimization framework for minimizing ischemia, the proposed system is capable of computing clamping points that reflect a deliberate and strategic surgical choice. This capability marks a clear transition from purely imitative or reactive behavior toward autonomous surgical reasoning. The robot is no longer merely following trajectories or reproducing human demonstrations; it is actively selecting the points of intervention that optimize clinical outcomes.

The results obtained across multiple patient anatomies demonstrate the robustness of the approach and highlight the potential of automated reasoning to support intraoperative decision-making. Importantly, the integration of this framework with a robotic execution pipeline shows that high-level surgical decisions can be translated into a coherent and safe sequence of actions performed autonomously by a robotic system. The proposed patient-to-model registration strategy and the method for computing collision-free clamping poses enable the robot to interact with the anatomy in a controlled and meaningful manner, closing the loop between planning and physical execution.

However, the system still relies on a critical assumption: the anatomical environment is static. All computations, from centerline extraction to ischemia estimation and pose generation, are performed on pre-operative data, and the robot executes the task under the assumption that the geometry of the kidney, arteries, and tumor remains unchanged throughout the procedure. In reality, surgical environments are inherently dynamic. Organs deform under the influence of gravity, respiration, and surgical manipulation. Arteries shift their positions, tissue tension changes, and physical interactions introduce a rigid and non-rigid variations that cannot be captured by a pre-operative model alone. Consequently, even a perfectly computed clamping point may not remain valid at the moment of robotic execution.

This exposes the central limitation of decision-making systems that operate without real-time perception: they lack contextual awareness. Surgical autonomy cannot be fully achieved unless the robot is capable of perceiving and adapting to changes in the operative field. The system must be able to update its internal model of the anatomy, revise plans when necessary, and adjust execution trajectories in response to evolving tissue states. Autonomous reasoning must therefore be tightly coupled with robust sensory understanding of the environment.

The work in this chapter thus lays the foundation for a new layer of autonomy: one in which high-level decisions are continuously informed by perceptual feedback. By defining and validating a complete pipeline for planning and executing selective clamping, we establish the reasoning component of the autonomy triad introduced at the beginning of the thesis. The next essential step is to integrate this reasoning with real-time deformation estimation, enabling the robot to track moving vessels, follow shifting tissue structures, and update the execution of clamping procedures as the anatomy evolves during surgery.

In summary, this chapter demonstrates that the robot can reason and plan, selecting optimal surgical strategies based on anatomical structure and ischemia minimization. Yet surgery unfolds in motion: organs deform, vessels shift, and contact dynamics evolve. To act safely and effectively, the robot must perceive, understand, and respond to these changes. This need for contextual awareness leads naturally to the next chapter, where deformation estimation and perception-driven adaptation are introduced as the final essential component for achieving intelligent and autonomous surgical behavior.

Chapter 4

Learning to Perceive: Understanding and Tracking Soft Tissue Deformation

4.1 Introduction: Percieving in Surgical Environment

One of the defining characteristics of surgical environments is the presence of highly deformable soft tissues. Kidneys, tumors, vessels, fascia, and surrounding structures all exhibit non-rigid behavior, continuously changing shape in response to gravity, physiological motion, and—most critically—the robot's own interactions. While the previous chapters of this thesis equipped the robot with the ability to imitate expert demonstrations and to reason about patient anatomy, these capabilities remain insufficient unless the robot can also understand how the surgical scene evolves during task execution. Autonomy in surgery requires not only the ability to act and decide but also the ability to perceive.

The objective of this chapter is to introduce the perceptual framework necessary for autonomous surgical robots to track, interpret, and respond to soft tissue deformation. Unlike rigid manipulation tasks where the geometry of the scene remains stable, surgical interventions require the robot to operate in an environment where every movement, every applied force, and every tool–tissue interaction may alter the anatomical configuration. A cutting path that was valid at the beginning of the procedure may deviate from its intended anatomical target seconds later. A target point on the tissue surface may shift due to tool contact or organ motion. If the robot continues to execute trajectories without awareness of these changes, even a learned or optimized policy may become unsafe or ineffective.

Therefore, understanding soft tissue deformation is not an optional enhancement but a fundamental prerequisite for safe and effective autonomous surgical execution. In this chapter, perception becomes the bridge between planning and action, allowing the robot to validate, update, and adapt its behavior based on real-time feedback from the surgical scene.

4.2 The Need for Understanding Soft Tissue Deformation

Accurate perception and estimation of deformation are essential for two key aspects of autonomous surgical execution: adaptive behavior and safety assurance. Soft tissues do not behave like rigid objects; their geometry evolves under even minimal interaction, and this evolution must be understood for safe and effective autonomous operation.

First, adaptive execution requires the robot to refine and update its trajectory based on how the anatomy is moving. The models learned in Chapter 2 assume that the tissue geometry is similar to that observed during demonstration. However, during real surgical tasks the tissue

rarely remains static. Pushing, grasping, cauterizing, or even maintaining contact produces local and global changes in shape. A trajectory that aligns with the tumor margin at time $t = 0$ must align with the *same anatomical margin* at time $t = 5$ seconds, not merely the same coordinates in space. This distinction is essential: autonomous execution must follow the anatomy, not the memory of a demonstration.

Second, deformation perception is fundamental for understanding how surgical plans are affected by anatomical motion. This is particularly evident when considering tasks such as selective clamping, where the computation of clamping points relies on the spatial configuration of arteries and their relationship to the kidney and tumor. While the planning framework introduced in Chapter 3 assumes a static pre-operative model, in actual surgical scenarios the kidney deforms significantly when manipulated, causing arterial branches to shift in space. Without an accurate model of these deformations, a clamping point that is optimal in the pre-operative geometry may no longer correspond to the correct arterial branch intraoperatively. Deformation awareness therefore becomes essential for maintaining the consistency between planned anatomical targets and their real-time locations, enabling future systems to revise or validate computed plans as the anatomy moves.

Third, perception of deformation provides the robot with the first mechanism for anomaly detection. If tissue begins to slip, deform excessively, or move in a way that deviates from the expected deformation pattern, a purely learned policy may continue execution blindly. In contrast, a perception-enabled system can detect discrepancies between expected and observed deformation, triggering adaptive strategies such as adjusting force, changing approach, or pausing the motion. Whereas imitation learning embeds implicit safety through expert-derived demonstrations, deformation tracking provides an explicit means for assessing whether the execution remains safe as the environment changes.

In summary, understanding soft tissue deformation is critical not only for enabling adaptive task execution but also for maintaining alignment between surgical intent and evolving anatomy. Although this chapter focuses primarily on tracking deformation in phantom experiments and evaluating how the robot adapts to rigid and non-rigid perturbations, the underlying perceptual principles lay the groundwork for enhancing more complex reasoning tasks, such as selective clamping, where anatomical accuracy and real-time consistency are essential.

4.2.1 Releted works and gaps in the literature

Robotic surgery has progressively evolved from purely teleoperated systems toward autonomous and semi-autonomous assistance, where a key requirement is the ability to understand, predict, and compensate for non-rigid organ deformation during surgical manipulation. This is particularly critical in minimally invasive procedures, such as robot-assisted partial nephrectomy (RAPN), where organ motions and deformations induced by respiration, tissue manipulation, and tool–tissue interaction significantly affect surgical precision and safety. Despite remarkable progress in computer vision, AR navigation, and deformable modeling, current systems still fail to provide the robot with a real-time, CT-consistent, complete representation of the organs under surgical manipulation. This work addresses this gap by introducing a tool-aware, non-learning, real-time deformation model applicable directly to preoperative CT meshes, requiring only image input, and enabling robotic guidance in deformable surgical scenes.

Deformable registration methods attempt to align preoperative CT-derived meshes or volumes with intraoperative endoscopic images. Several recent surveys highlight the importance of this task for AR and navigation. Han et al. [30] provide an extensive review of organ deformation modeling for AR navigation, emphasizing the need for reliable volumetric registration in liver surgery. Ali et al. [68] evaluate AR systems for laparoscopy, showing that deformable registration significantly reduces target registration error (TRE) compared to rigid overlays.

A large body of work focuses on using contours, shading, or partial surface reconstructions to drive CT-to-video alignment. Early works such as Koo et al. and Espinel et al. [69,70] utilize

monocular contour and shading cues for elastic registration. Espinel et al. [70] extend contour-based methods to multi-view settings using SfM/SLAM reconstructions for laparoscopic liver alignment, while Labrunie et al. [71] propose automatic rigid and deformable CT registration pipelines based on deep landmark detection and synthetic deformation data. Guan et al. [72] further incorporate stereo-reconstructed surfaces for deformable alignment in AR liver surgery. Heiselman et al. [73] review sparse intraoperative data registration strategies and discuss their limitations in complex deformation scenarios. Recent learning-based approaches, such as Pfeiffer et al., Cao et al., and Cao et al. [74–76] exploit synthetic FEM-generated data to train neural networks that infer deformation fields consistent with CT anatomy, achieving strong TRE performance but relying heavily on pre-computed training datasets and FEM-based simulation pipelines.

SLAM techniques reconstruct dense organ surfaces from video and track camera motion. Gao et al. survey the integration of SLAM into MIS, highlighting that most methods assume mild deformations or treat deformable surfaces with heuristic corrections. Classical deformable SLAM methods such as Song et al. [77, 78] use embedded deformation (ED) graphs to model soft-tissue motion from stereo vision, while Chen et al. [79] and Mahmoud et al. [80] reconstruct dense surfaces from monocular endoscopy. More recent works [81, 82]integrate geometric priors or neural fields for improved deformable mapping.

However, SLAM-based approaches suffer from three fundamental limitations for robotic guidance: they reconstruct only visible surfaces, lacking internal anatomy and complete geometry and they do not explicitly model tool-induced deformation, despite tools being the primary driver of organ motion, making them unsuitable for safety-critical robotic navigation. These limitations make SLAM-derived point clouds unsuitable as a canonical representation for robotic planning.

A parallel research direction focuses on estimating forces or interaction states directly from video. Giannarou et al. [83] recover brain tissue deformation for force estimation in neurosurgery. CNN- and transformer-based approaches [84–88] map local tissue deformation patterns to forces or stress. Pan et al. (2024) review such methods. These works confirm that tool–tissue interaction drives most deformation, but none produce a 3D deformation field on a CT mesh suitable for robotic control.

Several works explore control or AR updates under dynamic tissue. Zhan et al. [89] propose deformation-aware tissue scanning for ultrasound. Lagneau et al. [90] perform visual servoing on soft objects. Krupa et al. [91] demonstrate early semi-autonomous laparoscopic manipulation using visual servoing. More advanced works [92] integrate real-time deformation sensing for retinal surgery control, while Golse et al. [93] demonstrate dynamic AR overlays for liver surgery. Jiang et al. [94] discuss deformation-aware strategies in robotic ultrasound, and explore embodied intelligence for soft tissue manipulation.

Despite these advances, none provide a CT-mesh–consistent, tool-aware, purely image-driven non-rigid deformation field that can directly guide robot motion.

Although the literature provides important components—SLAM-based reconstruction, neural deformation estimators, FEM-based simulation, and AR overlays—several key challenges remain unaddressed:

As noted across methods [71, 74, 76] obtaining ground-truth organ deformation fields is extremely difficult. Most datasets are small, rely on synthetic FEM, or only capture surfaces—not full volumetric deformations. This is a major barrier to supervised learning.

As emphasized by reviews [30,36,68], real surgical data with synchronized CT meshes, surgical video, and deformation fields is scarce. Collecting such data is invasive, ethically constrained, and technically challenging.

Several approaches rely on FEM to generate training data or perform optimization. FEM provides biomechanical accuracy but is too slow for real-time robotic guidance, especially for large CT meshes.

NeRF and neural field methods produce high-fidelity surfaces but lack stable metric scaling,

deformation tracking, and CT alignment required for robotic manipulation.

As seen in force-estimation works, tools drive deformation, yet CT-registration and SLAM systems do not explicitly incorporate tool motion, despite its central importance during surgery.

For robot guidance and AR, only a CT mesh gives:full organ shape, semantic structure (tumours, vessels) and geometry, stable topology and therefore can be used as input for robot control and/or AR guidance for surgeons.

latly, many systems require FEM, synthetic datasets, force sensors or external markers. There is a need for a vision-only + CT mesh approach that works directly from intraoperative video.

To overcome these limitations, we propose the first tool-aware, non-learning, CT-mesh–consistent deformation model for robotic surgery, with the following contributions:

1. A vision-based deformation pipeline which integrates:
    - A vision-based non-learning deformation model: No training data, FEM simulation, or ground truth is required. Deformation arises from geometric warping defined directly on the CT mesh.
    - Tool-aware deformation from image cues: We explicitly track each tool independently and compute per-tool deformation fields, making the model sensitive to the true source of organ motion, together with organ deformation.
    - Rigid + non-rigid deformation integration: Our method jointly estimates organ pose and local deformations, avoiding drift associated with purely non-rigid models. Restoration (inverse deformation): We introduce a mechanism to recover the undeformed CT geometry for AR visualization or re-planning.
2. Integration with robot control or surgeon guidance: The deformation model is computationally light and can be integrated directly in the robot control loop or used as AR guidance application for surgeons.
3. Validated on diverse RAPN deformations: We test on multiple deformation scenarios in real RAPN videos, showing robust performance under large tool-induced motions.
4. Validation of the deformation-based control on silicone phantoms.

4.2.2 Position within the Autonomous Surgical Framework

Within the autonomy framework developed in this thesis, this chapter occupies a central and indispensable role. It extends the capabilities introduced in Chapters 2 and 3 by providing the perceptual foundation required for operating in realistic, deformable surgical environments.

Chapter 3, which focuses on surgical reasoning and the computation of optimal clamping points, relies entirely on a pre-operative model of the kidney and arteries. The planning framework introduced there assumes that the organ behaves as a rigid structure. However, in practice the kidney undergoes substantial deformation during surgical manipulation, causing anatomical structures, including the arteries targeted for clamping, to shift in space. Without a mechanism to perceive and track these changes, even the most carefully reasoned surgical plan risks becoming inaccurate the moment the anatomy deforms. Thus, the methods presented in this chapter are not merely a useful addition but a necessary extension for enabling Chapter 3 to operate robustly under realistic intraoperative conditions.

The relationship with Chapter 2 is more nuanced. The imitation learning framework developed there is trained on demonstrations recorded from deforming environments, particularly in the context of tumor cutting. As a result, deformation is implicitly embedded in the dataset and, to some extent, in the learned policy itself. However, such implicit exposure does not provide the robot with an explicit understanding of why the tissue is moving or how the deformation

influences the current surgical state. The robot can reproduce expert-like trajectories, but it lacks the ability to interpret deviations, analyze the underlying cause of tissue motion, or adapt its actions in response to unexpected deformations. Explicit deformation perception, therefore, complements imitation learning by transforming reactive imitation into proactive, context-aware execution.

Furthermore, the experimental framework developed in this chapter provides a natural extension of the tumor cutting task introduced in Chapter 2. While Chapter 2 demonstrated how a robot can imitate expert cutting trajectories based on visual and kinematic demonstrations, the experiments presented here allow us to investigate how the same task can be performed when the tissue undergoes deformation. By combining learned cutting trajectories with deformation tracking, we can evaluate whether the robot is capable of adapting the learned behavior to maintain alignment with the anatomical boundary as it moves. This constitutes a direct test of the robustness and adaptability of the learned models and provides insight into how deformation-aware perception can enhance the safety and precision of surgical tasks such as margin following, incision planning, and trajectory execution.

In this way, deformation perception serves as the connecting element that unifies the entire autonomy pipeline: imitation learning provides the skill, surgical reasoning provides the intent, and deformation tracking provides the awareness necessary to execute that intent safely and effectively in a dynamic anatomical environment. Rather than operating on fixed geometric assumptions, the robot can now observe how its actions alter the scene, interpret these changes, and adjust its motions accordingly. This chapter therefore marks the transition from static execution to dynamic adaptation, enabling all previous components of the autonomy framework to function reliably under real surgical conditions.

4.2.3 Testing Deformation Tracking in Realistic Phantom Experiments

To evaluate these perceptual capabilities, we focus on real-world experiments involving silicone kidney–tumor phantoms derived from patient CT scans. These models reproduce the mechanical and geometric characteristics of real soft tissues, providing a controlled yet realistic environment for studying tissue deformation and robotic adaptation.

The experiments in this chapter are not aimed at applying deformation awareness to specific clinical tasks such as clamping (addressed conceptually in later chapters), but rather at evaluating whether the robot can:

- track soft tissue deformation accurately in real phantoms,
- interpret how the deformation affects a precomputed trajectory,
- execute a surgical task (e.g., tumor cutting or margin following) under deformation,
- adapt its motion dynamically under both rigid and non-rigid perturbations.

Rigid perturbations correspond to global movements of the organ, mimicking patient motion or shifts in positioning. Non-rigid perturbations, induced by a secondary tool interacting with the tissue, simulate realistic surgical conditions where local deformation occurs near the region of interest. Evaluating the robot's ability to adapt in both cases is essential for validating the robustness of deformation-aware execution.

4.2.4 Structure of the Chapter

The remainder of this chapter is structured as follows. We first present the theoretical and computational background for modeling and estimating soft tissue deformation from visual data. We then describe the full deformation tracking pipeline, including both sparse and dense estimation methods. Next, we report the results of real-world experiments using silicone phantoms,

demonstrating how the robot tracks tissue motion and adapts its trajectory under rigid and non-rigid perturbations. Finally, we discuss the limitations of current sensing modalities and outline the path toward integrated multi-modal perception for autonomous surgical systems.

4.3 Methodological Framework

4.3.1 Overview

The proposed framework estimates non-rigid deformation of anatomical surfaces directly from stereo endoscopic video, relying solely on visual information and geometric reasoning. It maintains a consistent 3-D mesh representation of the organs, reconstructed from preoperative imaging, and continuously updates it according to observed motion and tool interaction. Rather than using data-driven priors or learned models, deformation is inferred analytically from depth, optical flow, and the spatial relationship between surgical instruments and tissue surfaces. The method integrates these cues through a unified formulation in which observed image motion is mapped to mesh vertices via barycentric rasterization, and local deformations are propagated across the surface using Gaussian-weighted diffusion. Elastic restoration and geometric regularization enforce temporal consistency and physical plausibility, allowing the organ to recover when tools are removed or motion ceases. The following subsections detail the mathematical formulation and computational stages of this vision-based non-rigid registration process. subsectionAlgorithm Summary

4.3.2 Non-rigid Deformation

Dense 3-D Flow From Optical Flow and Depth Variation

The 3-D displacement of a visible surface element is derived from the 2-D flow and the change in depth:

$$\Delta\mathbf{X}(x,y) = \begin{bmatrix} \frac{u(x,y)\, Z_t(x,y)}{f_x} \\ \frac{v(x,y)\, Z_t(x,y)}{f_y} \\ Z_{t+1}(x,y) - Z_t(x,y) ZoeDepth. \end{bmatrix}. \tag{4.1}$$

Equation (4.1) is evaluated for all pixels inside the organ mask; the resulting tensor $\mathbf{U}_t \in \mathbb{R}^{H\times W\times 3}$ constitutes the observed 3-D motion field.

Depth and 3-D Point Reconstruction

For each pixel (x, y) with depth $Z_t(x, y)$ we back-project into camera coordinates as

$$\mathbf{X}_t(x,y) = \begin{bmatrix} \frac{(x-c_x)}{f_x}\, Z_t(x,y) \\ \frac{(y-c_y)}{f_y}\, Z_t(x,y) \\ Z_t(x,y) \end{bmatrix}. \tag{4.2}$$

Pixels outside the binary organ mask $\mathcal{M}_{\text{mask}}$ are discarded. This yields a dense point cloud $\mathcal{P}_t = \{\mathbf{X}_t(x, y)\}$ representing the visible surface of the organ. A similar operation on the tool mask provides a tool point cloud $\mathcal{T}_t$.

Tool–Tissue Contact Detection

Let $\mathcal{V}_t = \{\mathbf{v}_i\}$ denote mesh vertices and $\mathcal{T}_t = \{\mathbf{t}_m\}$ the current tool points. A KD-tree query determines the minimum distance $d_{im} = \|\mathbf{v}_i - \mathbf{t}_m\|$ for each pair. A vertex is in contact if

$$\min_m d_{im} < \tau_c, \tag{4.3}$$

where τ_c (typically 1–2 mm) is the contact threshold. The indices of contacting vertices form the anchor set $\mathcal{A}_t$. This procedure is repeated independently for each instrument, yielding multiple disjoint anchor sets.

Rasterization and Barycentric Flow Accumulation

To map the dense 3-D flow of Eq. (4.1) from the image domain to the mesh domain, we render the mesh into the camera view using a differentiable rasterizer. For each pixel p that projects inside a triangle $f = (i_1, i_2, i_3) \in \mathcal{F}$ we obtain the triangle index and barycentric coordinates $\boldsymbol{\lambda}_p = (\lambda_1, \lambda_2, \lambda_3)$ satisfying $\sum_k \lambda_k = 1$.

Each pixel carries a 3-D flow vector $\Delta\mathbf{X}_p$. We distribute this flow to the triangle's vertices as

$$\Delta\mathbf{v}_{i_k}^{(p)} = \lambda_k^{(p)}\,\Delta\mathbf{X}_p, \quad k \in \{1,2,3\}. \tag{4.4}$$

A scatter-add operation then accumulates all contributions:

$$\mathbf{S}_i = \sum_{p:\, i \in f(p)} \lambda_i^{(p)}\,\Delta\mathbf{X}_p, \tag{4.5}$$

$$W_i = \sum_{p:\, i \in f(p)} \lambda_i^{(p)}, \tag{4.6}$$

$$\Delta\mathbf{v}_i = \frac{\mathbf{S}_i}{\max(W_i, \varepsilon)}. \tag{4.7}$$

Equations (4.4)–(4.7) produce the per-vertex displacement field consistent with visible motion.

Propagation to Non-Visible Regions

Pixels that are occluded or outside the field of view have no direct observation. To ensure smooth tissue motion, we diffuse the displacement from anchor vertices $\mathcal{A}_t$ or visible vertices to all other vertices through a Gaussian kernel on the mesh surface:

$$\Delta\mathbf{v}_i = \eta \sum_{a \in \mathcal{N}_k(i)} \frac{\exp\!\big(-\|\mathbf{v}_i - \mathbf{v}_a\|^2/2\sigma^2\big)}{\sum_{a'} \exp\!\big(-\|\mathbf{v}_i - \mathbf{v}_{a'}\|^2/2\sigma^2\big)}\,\mathbf{f}_a, \tag{4.8}$$

where $\mathbf{f}_a$ is the flow at anchor a, σ controls spatial influence, k is the number of neighbors, and η a magnitude scale factor. A global semi-rigid component is added by blending with the mean anchor flow $\bar{\mathbf{f}}$:

$$\Delta\mathbf{v}_i^{\text{final}} = (1-\gamma_i)\Delta\mathbf{v}_i + \gamma_i\bar{\mathbf{f}}, \qquad \gamma_i = \exp\!\Big(-\frac{d_i^2}{2\sigma_g^2}\Big), \tag{4.9}$$

where d_i is the distance from vertex i to the nearest anchor and σ_g controls the extent of global motion.

Elastic Restoration

When the tool retracts or reverses direction, the tissue gradually returns toward its rest configuration. Let $\mathbf{d}_{t-1}$ and $\mathbf{d}_t$ denote consecutive tool direction vectors. If the inversion angle $\theta = \cos^{-1}\big(\frac{\mathbf{d}_{t-1}\mathbf{d}_t}{\|\mathbf{d}_{t-1}\|\|\mathbf{d}_t\|}\big)$ exceeds a threshold θ_{inv} (typically 135°), restoration is activated with a gain

$$\kappa = \kappa_0 \frac{\theta - \theta_{\text{inv}}}{\pi - \theta_{\text{inv}}}. \tag{4.10}$$

Each previously deformed vertex $i \in \mathcal{R}_t$ is updated as

$$\mathbf{v}_i \leftarrow \mathbf{v}_i + \kappa\,(\mathbf{v}_i^{(0)} - \mathbf{v}_i), \tag{4.11}$$

where $\mathbf{v}_i^{(0)}$ are the rest-shape vertices. This process is performed independently for each instrument, enabling concurrent multi-tool restoration.

Mesh Regularization and Taubin Smoothing

To remove local irregularities while preserving curvature, we perform Laplacian smoothing over the subset of modified vertices $\mathcal{V}_\Delta$:

$$\mathbf{v}_i^{\text{new}} = (1-\alpha)\mathbf{v}_i + \alpha \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{v}_j, \tag{4.12}$$

where $\mathcal{N}(i)$ is the 1-ring neighborhood of i and $\alpha \in [0,1]$ controls smoothing strength.

Cutting Path Definition and Deformation

In addition to estimating organ deformation, we maintain a deformation-aware representation of the surgical cutting path.

The resection path is defined directly from the preoperative CT meshes by leveraging anatomical and oncological principles. In partial nephrectomy, the recommended safe surgical margin around the tumor ranges from 1–3 mm depending on tumor grade and location. To reflect this clinical guideline, we construct the initial cutting path $\mathcal{C}_0$ as the intersection between the kidney surface $\partial\Omega_{\text{kidney}}$ and the tumor surface expanded by a fixed 2 mm offset:

$$\mathcal{C}_0 = \partial\Omega_{\text{kidney}} \cap \big(\partial\Omega_{\text{tumor}} \oplus 2\,\text{mm}\big), \tag{4.13}$$

where $\oplus$ denotes a Minkowski sum with a sphere of radius 2 mm. This provides a geometrically well-defined and medically justified resection boundary that achieves tumor removal while preserving healthy parenchyma.

During surgical manipulation, the organ undergoes non-rigid deformation. To maintain a physically plausible and anatomically correct trajectory, the cutting path must deform coherently with the surrounding tissue. We therefore advect each path point according to the same vertex displacement field used for mesh deformation. Let $\mathbf{c}_j^t$ be the j-th path point at frame t, and let $\Delta\mathbf{v}(\mathbf{x})$ denote the estimated displacement at a surface location $\mathbf{x}$. The updated cutting path is computed as

$$\mathbf{c}_j^{t+1} = \mathbf{c}_j^t + \Delta\mathbf{v}(\mathbf{c}_j^t), \tag{4.14}$$

which preserves continuity, smoothness, and local geometric consistency. As a result, the cutting path remains anatomically meaningful and visually aligned with the evolving organ shape. This dynamic update serves two complementary purposes. First, it provides an interpretable and visually verifiable indicator of deformation accuracy: if the mesh deformation is correct, the projected cutting path should remain aligned with the surgical tool movements visible in the video. Second, the evolving path is used later in this work as the reference trajectory for robot execution in phantom experiments, where the robot follows the deformation-aware cutting curve as the tissue undergoes manipulation. By coupling the cutting path to the deformation model, the framework preserves anatomical correctness while enabling both visual validation and downstream robotic control.

Algorithm 2 Vision-Based Non-Rigid Registration

Input: rectified stereo frames I_t^L, I_t^R, calibration $\mathbf{K}$, meshes $\mathcal{M}_{t-1}$.
Compute depth Z_t and segmentation masks.
Reconstruct organ and tool point clouds using Eq. (4.2).
Estimate 3-D flow $\mathbf{U}_t$ via Eq. (4.1).
Rasterize $\mathcal{M}_{t-1}$ and accumulate per-pixel flows using Eqs. (4.4)–(4.7).
Detect tool–tissue contacts using Eq. (4.3).
Propagate displacements using Eq. (4.8) and blend via Eq. (4.9).
Apply elastic restoration if $\theta > \theta_{\text{inv}}$ using Eq. (4.11).
Smooth the mesh via Eq. (4.12).
Update the cutting path $\mathcal{C}_\delta$ accordingly.
Output: updated mesh $\mathcal{M}_t$ and path $\mathcal{C}_\delta$. =0

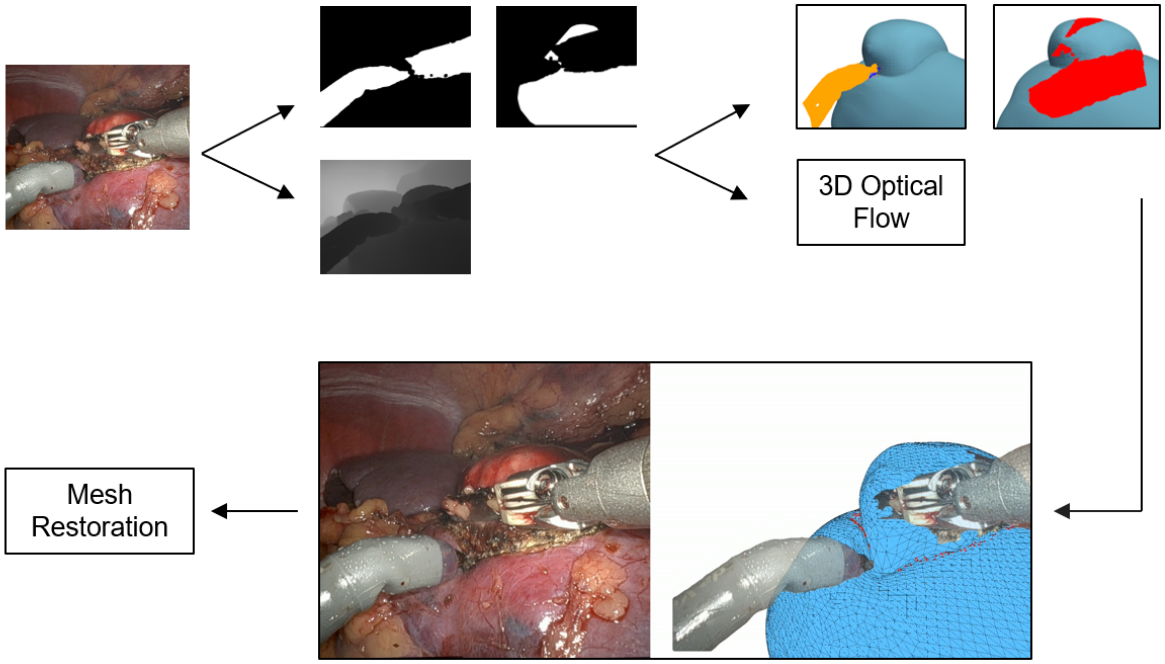


Figure 4.1: Proposed deformation pipeline: stereo video → segmentation and depth → tool–organ interaction → deformation transfer → optical flow → updated organs mesh and cutting path → mesh restoration.

4.3.3 Rigid Registration

Non-rigid deformation is not enough for the correct and accurate execution of a task involving deformable and moving object as organs: when the tissue is pushed or shifted by the surgical tools, the global pose of the organ changes in the camera frame. Since the deformable model is defined in the coordinate system of the preoperative CT mesh, we must continuously track and update the global transformation that keeps the mesh in correspondence with the observed scene. Without this step, the mesh would drift out of the image plane, and subsequent non-rigid updates would no longer reflect the true anatomy.

To accomplish this alignment, we employ a frame-by-frame rigid registration implementing a point-to-point GPU-based ICP variant. The method follows the classical structure of ICP but introduces three differences relative to standard CPU-based implementations: (i) corre-

spondences are computed through fully parallel GPU k-nearest-neighbor queries, (ii) the rigid transform is estimated in a batched fashion using a closed-form SVD, and (iii) the cumulative rigid motion is explicitly maintained for direct application to the organ mesh at each iteration.

Let the source and target point clouds be

$$\mathbf{X}^{(0)} \in \mathbb{R}^{B\times N\times 3}, \qquad \mathbf{Y} \in \mathbb{R}^{B\times M\times 3},$$

where B is the batch size (e.g., multiple frame pairs), N and M are the number of points in source and target clouds, respectively. At iteration k, ICP proceeds in two steps.

Correspondence Search (GPU k-NN) We first compute, for each transformed source point $\mathbf{x}_i^{(k)}$, its closest target point using a GPU-based k-nearest-neighbor query:

$$\mathbf{y}_i^{(k)} = \arg\min_{\mathbf{y}\in\mathbf{Y}} \|\mathbf{x}_i^{(k)} - \mathbf{y}\|_2,$$

This replaces the CPU KD-tree typically used in standard ICP, and enables parallel correspondence computation for all batches and points.

SVD-based Rigid Transform (Batched) Given the correspondences $\{(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(k)})\}$, we estimate the optimal rigid transform in closed form for each batch element. For batch b, we compute the centroids

$$\bar{\mathbf{x}}_b^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{b,i}^{(k)}, \qquad \bar{\mathbf{y}}_b^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_{b,i}^{(k)},$$

and form the centered point sets

$$\tilde{\mathbf{x}}_{b,i}^{(k)} = \mathbf{x}_{b,i}^{(k)} - \bar{\mathbf{x}}_b^{(k)}, \qquad \tilde{\mathbf{y}}_{b,i}^{(k)} = \mathbf{y}_{b,i}^{(k)} - \bar{\mathbf{y}}_b^{(k)}.$$

The cross-covariance matrix is

$$\mathbf{H}_b^{(k)} = \sum_{i=1}^{N} \tilde{\mathbf{x}}_{b,i}^{(k)} \big(\tilde{\mathbf{y}}_{b,i}^{(k)}\big)^\top \in \mathbb{R}^{3\times 3},$$

and its singular value decomposition,

$$\mathbf{H}_b^{(k)} = \mathbf{U}_b^{(k)}\, \mathbf{\Sigma}_b^{(k)} \big(\mathbf{V}_b^{(k)}\big)^\top,$$

yields the batch-wise rotation

$$\mathbf{R}_b^{(k)} = \mathbf{V}_b^{(k)} \big(\mathbf{U}_b^{(k)}\big)^\top.$$

To avoid reflections, we enforce $\det(\mathbf{R}_b^{(k)}) = 1$ by flipping the last column of $\mathbf{V}_b^{(k)}$ when necessary. The corresponding translation is

$$\mathbf{t}_b^{(k)} = \bar{\mathbf{y}}_b^{(k)} - \mathbf{R}_b^{(k)} \bar{\mathbf{x}}_b^{(k)}.$$

The transform is then applied to all source points in batch b:

$$\mathbf{x}_{b,i}^{(k+1)} = \mathbf{R}_b^{(k)}\, \mathbf{x}_{b,i}^{(k)} + \mathbf{t}_b^{(k)}.$$

Unlike many standard ICP implementations, we explicitly maintain the cumulative transform

$$\mathbf{R}_b^{\text{all}} \leftarrow \mathbf{R}_b^{(k)} \mathbf{R}_b^{\text{all}}, \qquad \mathbf{t}_b^{\text{all}} \leftarrow \mathbf{R}_b^{(k)} \mathbf{t}_b^{\text{all}} + \mathbf{t}_b^{(k)},$$

so that $\mathbf{R}_b^{\text{all}}, \mathbf{t}_b^{\text{all}}$ represent the final rigid registration from the original source to the target.

Convergence Criterion The mean squared correspondence error

$$E^{(k)} = \frac{1}{BN} \sum_{b=1}^{B} \sum_{i=1}^{N} \left\| \mathbf{x}_{b,i}^{(k)} - \mathbf{y}_{b,i}^{(k)} \right\|_2^2$$

is monitored at each iteration. The optimization terminates when the change $|E^{(k)} - E^{(k-1)}|$ falls below a tolerance, or when a maximum number of iterations is reached. This simple, distance-based stopping rule is compatible with the batched GPU formulation and avoids additional bookkeeping.

4.4 Evaluation Metrics

Evaluating non-rigid deformation in surgical scenes is challenging because no ground-truth 3-D trajectories or vertex displacements are available. Instead of measuring absolute geometric accuracy, we quantify how well the deformed mesh remains consistent with the observations that generated the deformation: image contours, optical flow, photometric appearance, and depth. These measures evaluate whether the estimated deformation explains the visual evidence, which is the only feasible form of supervision in real surgical datasets.

4.4.1 Silhouette Consistency

A correctly deformed mesh must reproduce the organ's visible contour when rendered into the camera frame. We project the mesh at time t and compute its binary silhouette mask $\mathcal{S}_t^{\text{mesh}}$. This is compared against the segmentation mask obtained from the image, $\mathcal{S}_t^{\text{img}}$, using an L_1 discrepancy:

$$E_{\text{sil}} = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} \left| \mathcal{S}_t^{\text{mesh}}(\mathbf{p}) - \mathcal{S}_t^{\text{img}}(\mathbf{p}) \right|. \tag{4.15}$$

This metric captures global misalignments and large-scale deformation errors. The average silhouette discrepancy across the entire sequence was 1×10^{-2}, indicating that the mesh projection remains visually overlapped on the observed organ outline.

4.4.2 Flow Consistency

Mesh deformation should generate motion consistent with the optical flow observed in the RGB sequence. To measure this, the predicted 3-D displacement of each visible vertex is reprojected to the image plane, yielding a synthetic flow field $\mathbf{u}_t^{\text{mesh}}$. This is compared to the dense optical flow computed directly from the RGB images, $\mathbf{u}_t^{\text{img}}$:

$$E_{\text{flow}} = \frac{1}{|\Omega_{\text{vis}}|} \sum_{\mathbf{p} \in \Omega_{\text{vis}}} \left\| \mathbf{u}_t^{\text{mesh}}(\mathbf{p}) - \mathbf{u}_t^{\text{img}}(\mathbf{p}) \right\|_2. \tag{4.16}$$

The resulting average error was below 0.5 pixels, demonstrating strong temporal coherence between the estimated mesh deformation and true image motion.

4.4.3 Depth Consistency

A valid mesh should also reproduce the stereo-derived depth map. For each visible pixel we project the deformed mesh and compute the discrepancy between predicted depth Z_t^{mesh} and stereo reconstruction Z_t^{stereo}:

$$E_{\text{depth}} = \frac{1}{|\Omega_{\text{vis}}|} \sum_{\mathbf{p} \in \Omega_{\text{vis}}} \left| Z_t^{\text{mesh}}(\mathbf{p}) - Z_t^{\text{stereo}}(\mathbf{p}) \right|. \tag{4.17}$$

The mean absolute depth deviation was approximately 2 mm, which is within the noise range of stereo reconstruction and indicates accurate global placement of the mesh with respect to the true surface.

4.4.4 Photometric Consistency

Because tissue is observed under near-constant lighting over short temporal windows, the appearance of corresponding surface points should remain similar after deformation. For each vertex visible at times t and $t+1$, we compare its projected pixel intensity:

$$E_{\text{photo}} = \frac{1}{|\Omega_{\text{vis}}|} \sum_{\mathbf{p} \in \Omega_{\text{vis}}} \left\| I_t(\mathbf{p}) - I_{t+1}(\mathbf{p} + \mathbf{u}_t^{\text{mesh}}(\mathbf{p})) \right\|_1 . \tag{4.18}$$

This photometric error validates whether the mesh-derived motion correctly warps the image to match the next frame. In our experiments the mean photometric discrepancy remained below 5% of the intensity dynamic range, confirming that the deformation explains the frame-to-frame color changes.

4.4.5 Cutting Path Consistency

Although the cutting path is defined anatomically in 3-D, its correctness in real surgical video must be evaluated in the observable domain. Because the surgical tool trajectory is directly measurable only in the image plane and no ground-truth 3-D cutting trajectory exists, we compute a Cutting Path Consistency (CPC) error that assesses visual agreement between the projected cutting path and the tool motion.

Let $\mathcal{C}_t$ be the deformation-aware cutting path at time t, and let Γ_t be the tool trajectory extracted from the image. We project each 3-D path point into the camera frame via the calibrated projection $\pi(\cdot)$:

$$\mathcal{C}_t^{\text{vis}} = \{\pi(\mathbf{c}) \mid \mathbf{c} \in \mathcal{C}_t\}.$$

Because parts of the cutting path may be occluded by instruments or tissue folds, we evaluate consistency only on the visible subset $\mathcal{C}_t^{\text{vis}}$, determined through rasterization using a Z-buffer visibility test. The CPC error is defined as

$$E_{\text{CPC}}(t) = \frac{1}{|\mathcal{C}_t^{\text{vis}}|} \sum_{\mathbf{c} \in \mathcal{C}_t^{\text{vis}}} \min_{\mathbf{g} \in \Gamma_t} \|\pi(\mathbf{c}) - \mathbf{g}\|_2 . \tag{4.19}$$

This metric captures functional visual correctness: a low CPC error indicates that the deformation-aware path evolves coherently with the actual tool motion visible in the video. Pixel units are used because the tool trajectory is reliably observable only in 2-D.

The average CPC error was approximately 5 pixels (less than 1% of the image width), demonstrating strong agreement between the updated cutting path and the visible surgical tool motion. This validates both the physical plausibility of the deformation model and the visual consistency of the resulting resection trajectory. In later phantom experiments, we will additionally report physical path-tracking errors in centimeters.

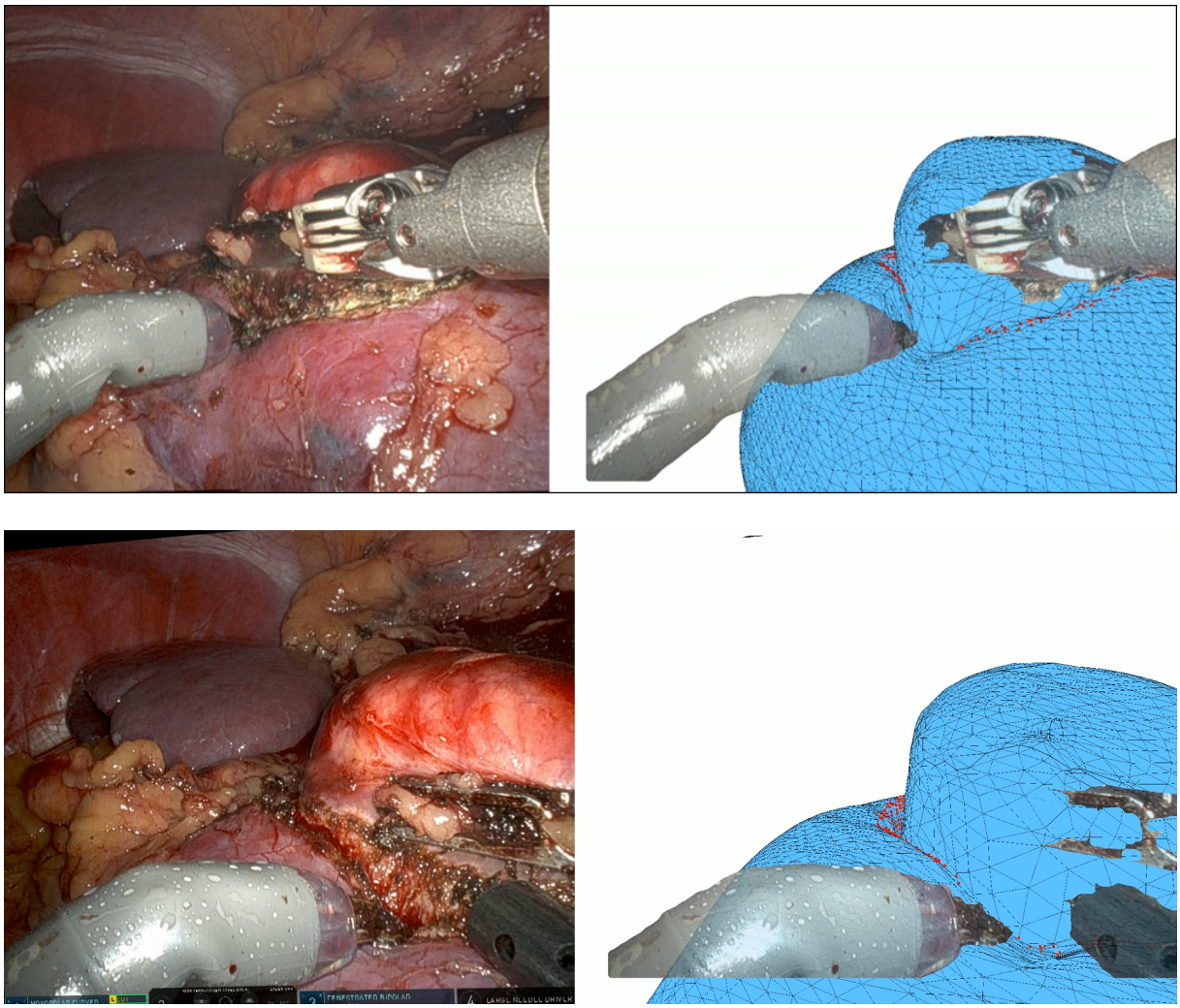


Figure 4.2: Examples of soft-tissue deformation estimated from surgical video sequences. The deformation pipeline successfully captures both organ motion and tool–tissue interactions, producing coherent displacement fields across the kidney surface. The red points represent the surgical cutting path, which is continuously updated according to the estimated deformation field, preserving geometric consistency throughout the procedure.

4.5 Conclusion: Percieving modalities and future directions

This chapter set out to address a central challenge in autonomous and semi-autonomous surgical robotics: the ability to perceive, quantify, and track non-rigid soft-tissue deformation in real time using only visual information. Surgical environments are inherently dynamic. Organs shift due to respiration, heartbeat, tool manipulation, tissue tension, and retraction. For a robot to execute a safe and effective plan—particularly in procedures such as robot-assisted partial nephrectomy (RAPN)—it must continuously maintain an updated understanding of how the anatomy is deforming. The work presented in this chapter tackled this problem by developing a computational pipeline capable of estimating the full, time-varying deformation field directly from video.

The proposed system leverages robust visual feature tracking combined with geometric alignment and registration procedures to generate a dense and temporally consistent map of tissue motion. At each frame, tracked points are propagated forward and backward in time, filtered, and embedded within a continuous geometric model to yield a spatially smooth deformation

field. This field is then evaluated at arbitrary query points to provide displacement information relevant to robotic actions. Importantly, the approach achieves low latency by restricting computation to lightweight, image-based operations and by avoiding heavy biomechanical solvers such as FEM or learning-based architectures requiring GPU inference. This aligns directly with the real-time constraints of intraoperative use.

The pipeline was validated in early physical phantom experiments, where displacements and relative motion patterns could be measured indirectly. Across these validations, the system demonstrated promising accuracy: low Target Registration Error (TRE), consistent point-tracking accuracy, and reliable displacement estimation under varying camera poses and deformation types. These results establish the system's ability to generalize beyond a single controlled scenario, capturing the essential deformation phenomena that occur in soft-tissue surgery.

The principal contribution of the chapter is therefore the design and implementation of a perceptually-aware soft-tissue tracking system that delivers time-dependent spatial information to the rest of the autonomy pipeline. It provides the robot with a real-time estimate of how the environment is evolving, forming the perceptual foundation required for safe, adaptive behaviour.

4.5.1 The Pivotal Role of Perceiving in the LRPA Paradigm

Within the broader LRPA (Learning–Reasoning–Perceiving–Acting) framework that guides this thesis, the work presented here fills the critical role of Perceiving. While Learning provides a structured understanding of the task and Reasoning determines how best to adapt the high-level plan to new contexts, neither component can operate safely without up-to-date information about the actual intraoperative state. In soft-tissue surgery, this state is not static but undergoes continuous, richly non-rigid deviations from the preoperative shape. Thus, Perceiving is responsible for detecting, quantifying, and interpreting these deviations, producing a corrected representation

A key argument of this chapter is that autonomous systems are fundamentally limited without accurate perception of non-rigid motion. In the LRPA pipeline, Perceiving ensures:

(1) Validity of Plans. Even the most carefully learned trajectory becomes obsolete if the target anatomy undergoes displacements, a common occurrence in RAPN. The presented system corrects the spatial substrate on which the plan operates, ensuring that a pre-learned policy remains physically meaningful.

(2) Support for Reasoning. Reasoning modules require grounded, geometry-aware representations of the surgical field. By providing a continuously updated deformation field, Perceiving enables the robot to evaluate whether a clamping point must shift, whether a cutting trajectory must be reshaped, or whether the anatomy has moved outside safe bounds. Reasoning thus becomes tightly coupled to the real scene rather than abstracted preoperative models.

(3) Enhanced Safety. Non-rigid deformation is often a precursor to task failure or hazardous conditions. Rapid, unexpected movement may indicate excessive tool forces, loss of tension, or unintended collisions. Perceiving acts as a first layer of anomaly detection: it identifies deformation patterns incompatible with expected behaviour and can trigger immediate control adjustments or safety halts.

In essence, Perceiving forms the bridge between the real world and the robot's internal state representation. It prevents kinematic blindness, enabling autonomy that is truly context-aware.

4.5.2 Limitations and Areas for Future steps

Despite the achievements presented, several limitations remain. These are not weaknesses but rather opportunities for deeper, more sophisticated follow-up work.

The system depends exclusively on visual cues. While this feature reamains solid point of the work, that can also creates constraints on performance in the presence of motion blur, illumination changes, occlusions (especially from tools or blood), texture-poor surfaces, and

smoke. Moreover, purely visual tracking lacks the ability to sense force-based interactions, tissue stiffness, or mechanical response—factors crucial in predicting how tissue will deform under load.

Validation so far used simple manipulation primitives (e.g., pushing with a passive probe). To understand deformation under more realistic loading—clamping, grasping, sweeping—future work must incorporate an actuated surgical instrument capable of applying controlled forces while simultaneously being tracked. Such a tool is already under development and will enable richer experiments.

While the method is robust, it still requires further generalization across different lighting and white-balance conditions, organ shapes and appearances, camera viewpoints and optics, and dynamic patterns unique to different procedures.

Self-Supervised Learning (SSL) can be introduced to adapt the perception pipeline to unlabeled surgical videos at scale, dramatically improving performance without requiring ground-truth annotations.

The current pipeline is computationally efficient but not as fast as potential deep learning alternatives. An important insight is that the system itself could serve as a pseudo-ground-truth generator, providing supervised targets for training CNN or Transformer-based displacement estimators. This would enable low-latency inference suitable for embedding directly into the control loop.

The environments tested so far are foundational. To ensure clinical applicability, the pipeline must be validated on more complex tasks resembling real surgical manipulations, across multiple anatomies. These future experiments will challenge the system under conditions approximating full surgical complexity.

4.6 Future Steps in Perception Roadmap

Building on the foundations established here, several strategic directions emerge for advancing the Perceiving module and enhancing the autonomy stack.

The imminent availability of an actuated, force-sensing tool will introduce mechanical data that can be fused with visual deformation. From this fusion, the robot could infer tissue compliance, local stiffness, and damping characteristics, producing richer deformation predictions and enabling reasoning grounded in biophysical insight. In parallel, optional multi-modal data sources (e.g., ultrasound, structured light) could complement the visual pipeline where depth ambiguity and occlusion hinder performance.

The deformation fields computed in this chapter can be distilled into a training set for a deep learning surrogate model. Such a model could predict deformation directly from raw video in a fraction of the time required by geometry-based methods. This would reduce system latency and open the door to higher-frequency control.

SSL techniques will further allow the model to generalize across diverse surgical scenes without requiring manual labels, leveraging the growing availability of clinical video datasets.

To demonstrate broad utility, the system should be validated across a range of surgical tasks (suturing, dissection, retraction) and organ types (kidney, liver, bowel). Each of these tissues exhibits unique deformation characteristics that will stress-test the perception pipeline and encourage novel algorithmic extensions.

Finally, ongoing work will integrate the deformation signal more tightly with the Reasoning component, refining downstream modules for trajectory correction, task monitoring, and safe motion planning. The long-term vision is a fully adaptive surgical robot in which Perceiving continuously informs Reasoning, enabling Actions that remain precise even under substantial anatomical shifts

Chapter 5

From Awareness to Automation: Integrating Learning, Reasoning, and Perception

5.1 Introduction: Bridging Perception, Reasoning, and Action

The previous chapters explored, in a progressive manner, the foundations of surgical autonomy. First, imitation learning allowed the robot to acquire expert surgical behaviors through demonstration, enabling it to act and decide within well-defined scenarios. Then, reasoning and decision-making mechanisms, exemplified by the selective clamping framework, empowered the system to autonomously plan surgical actions based on anatomical understanding. Subsequently, perception and deformation modeling closed the sensory gap by allowing the robot to perceive the continuous changes of soft tissues and to adapt its behavior to the dynamic surgical scene.

In this chapter, these three layers, learning, reasoning, and perception, are integrated into a unified framework. Perception of soft-tissue deformation, obtained from the methods described in the previous chapter, is now exploited to dynamically update surgical strategies and execution trajectories. This capability represents a crucial step toward context-aware surgical automation, where robotic actions are not only pre-planned or learned but also continuously informed by real-time perceptual feedback.

We demonstrate this integration in two representative applications:

1. *Selective clamping with deformation-aware artery tracking*, where the system uses deformation information to maintain the alignment of clamping points as the organ deforms.

2. *Tumor cutting and trajectory adaptation under deformation*, where the robot follows a precomputed cutting path while adapting to deformation perturbations generated by tool interactions.

Finally, we present the conceptual and mechanical design of a new surgical tool that can be mounted on a robotic arm and automatically controlled, forming the physical interface between reasoning, perception, and action.

5.2 Integration of Deformation Awareness into Surgical Reasoning

The ability to reason about the surgical anatomy and plan high-level actions is a crucial component of autonomous robotic surgery. However, as shown in the previous chapters, both imitation learning and surgical reasoning rely, either explicitly or implicitly, on assumptions about the stability of the anatomical environment. In reality, surgical scenes are dominated by soft tissues

that deform continuously under the influence of gravity, respiration, fluid pressure, and, most importantly, tool–tissue interaction. A surgical plan computed on a pre-operative model is therefore only an initial snapshot of the anatomical configuration. As soon as the organ is manipulated, the geometry changes, requiring the robot to update its internal understanding of the anatomy to maintain the validity and safety of its actions.

This chapter introduces the integration of deformation awareness with surgical reasoning, forming a key connection between perception and high-level decision-making. The deformation tracking framework developed in the previous chapter offers a real-time, continuous estimate of how each point of the organ surface moves in response to external forces. This estimate is captured in the deformation field $\mathbf{D}(\mathbf{x}, t)$, a function that maps each point $\mathbf{x}$ from an initial reference configuration to its displaced location at time t. Using this formulation, any static reference point $\mathbf{p}_0$ can be updated to its real-time location through the relation

$$\mathbf{p}(t) = \mathbf{p}_0 + \mathbf{D}(\mathbf{p}_0, t), \tag{5.1}$$

which provides the foundational mechanism for propagating anatomical landmarks, target points, trajectories, or planning-relevant structures into the current deformed state of the organ. This update rule is the perceptual bridge between planning and execution: it ensures that strategies computed in the reference model remain aligned with the dynamically evolving anatomical reality.

In this chapter, we focus on how this perceptual awareness can be integrated with surgical reasoning and, in particular, how it enhances the robustness and accuracy of selective vascular clamping. While the primary goal here is conceptual—demonstrating how deformation awareness strengthens high-level planning—the benefits extend to many surgical tasks in which anatomy changes as a consequence of robotic manipulation.

5.3 Deformation-Aware Selective Clamping: Motivation and Framework

5.3.1 Introduction and motivations

Selective clamping represents a compelling application of deformation-aware reasoning because it relies heavily on precise spatial understanding of the arterial tree. The computation of clamping points introduced in the previous chapter assumes a static, pre-operative configuration of the kidney and its vasculature. However, soft tissue motion and deformation pose a major challenge to this assumption. Kidney mobilization, instrument pressure, respiratory motion, and tissue tension can shift the position of arterial branches significantly. Even small displacements can render a pre-planned clamping point inaccurate, potentially leading to ineffective occlusion or unintended manipulation of the wrong branch.

To address this limitation, we extend the selective clamping framework by integrating the deformation field $\mathbf{D}(\mathbf{x}, t)$ into the planning stage. The enhanced procedure begins with a manual alignment between the 3D anatomical models (kidney, tumor, arterial tree) and the intraoperative image frame. This establishes the reference state from which all subsequent updates are computed. Once aligned, each anatomically relevant structure—such as the artery centerline, bifurcation points, or clamping targets—is treated as a reference point $\mathbf{p}_0$ or set of points to be propagated through Equation 5.1. The deformation field then updates the position of these structures as the kidney deforms, allowing the computed clamping points to follow their corresponding arterial branches throughout the procedure.

This capability provides a major safety advantage. Because the position of the arteries is continuously updated, the robot or surgeon always operates using clamping locations consistent with the current anatomy, not the initial model. This significantly reduces the risk of misalign-

ment, vessel injury, or partial occlusion, and ensures that the surgical plan remains optimally aligned with the real-time vascular geometry.

5.3.2 The Dynamics of Arterial Deformation

Arteries exhibit unique deformation characteristics that require special treatment compared to general soft tissues. They are slender tubular structures whose radius is small relative to their length, and their behavior under deformation is governed both by local interactions and by global motions transmitted through the surrounding organ. To provide a precise description of arterial deformation, we introduce a three-tiered hierarchy that distinguishes between different scales and modes of deformation.

1. Very Localized Micro-Deformation (Transcurable). At the most local scale, arteries may experience small surface compressions or minute displacements caused by light contact or pressure. These deformations are often localized to tiny regions and do not meaningfully alter the trajectory or topology of the centerline. For selective clamping, these micro-deformations are *transcurable*: as long as the global position of the arterial branch remains stable, micro-level distortions do not meaningfully affect the selection or placement of clamping points.

2. Significant Local Branch Motion (Tool-Induced). A more substantial form of deformation occurs when surgical tools interact with the vasculature directly or indirectly. For example, a grasper may lift or retract an arterial branch to expose the target vessel. In such cases, the deformation is locally non-rigid and concentrated in one portion of the arterial tree. The affected branch may rotate, bend, or translate relative to its original configuration, while nearby branches may remain unchanged. This type of motion has a major impact on clamping because the local geometry of the branch—and therefore the spatial location of the clamping point—can change dramatically. Tracking deformation at this level ensures that selected clamping targets remain valid even as individual branches are manipulated.

3. Wider Organ-Coupled Deformation (Global). The most pervasive form of deformation arises from motion of the organ as a whole. When the kidney shifts due to grasping, rotation, or physiological forces, the entire arterial tree experiences a combination of rigid displacement and non-rigid stretching or bending. This wide deformation affects every clamping point and branch simultaneously, making it essential to propagate the entire arterial structure through the deformation field. Capturing these global changes ensures that the clamping strategy remains coherent with the overall anatomical context.

5.3.3 Implementation of Deformation-Aware Selective Clamping

The integration of deformation awareness into selective clamping requires a precise and continuous coupling between the pre-operative vascular model and the intraoperative anatomical configuration. The implementation pipeline described here translates the deformation field $\mathbf{D}(\mathbf{x}, t)$ into real-time updates of arterial geometry and clamping targets, enabling the robot to operate on anatomically accurate and dynamically consistent information.

Perception-to-Action Pipeline

The process begins with two essential inputs. First, the pre-operative 3D vascular model provides the reference configuration, including the artery centerlines, bifurcations, and the precomputed clamping point $\mathbf{p}_c$ located on the reference centerline. Second, intraoperative image data—typically stereo or RGB-D frames—serve as the input to the deformation estimation pipeline introduced in the previous chapter. This pipeline produces a dense deformation field

$\mathbf{D}(\mathbf{x}, t)$ that assigns a displacement vector to each point $\mathbf{x}$ in the scene, capturing both local and global tissue motion.

At each time step, the deformation field is applied to the entire vascular model. Because arteries are represented as 1D tubular structures, the centerline becomes the key geometric feature for propagation. Each point $\mathbf{p}_0$ on the reference centerline is updated by sampling the deformation field at that location and computing

$$\mathbf{p}(t) = \mathbf{p}_0 + \mathbf{D}(\mathbf{p}_0, t). \tag{5.2}$$

This operation yields the deformed centerline at time t, reflecting the motion of the arterial branch as the organ deforms. The clamping point $\mathbf{p}_c$, originally defined on the reference centerline, undergoes the same update:

$$\mathbf{p}_c(t) = \mathbf{p}_c + \mathbf{D}(\mathbf{p}_c, t), \tag{5.3}$$

thereby ensuring that the clamping target remains physically attached to its underlying anatomical structure. In this way, the system continuously tracks the precise location of the artery segment to be occluded, maintaining the spatial consistency between the planned surgical strategy and the real-time tissue geometry.

Diameter Awareness for Safe Clamping

While accurate centerline tracking provides the correct 3D position for clamping, it does not directly specify how the robot should perform the occlusion. Clamping is an action, not merely a positioning task, and the aperture and applied force of the tool must be adapted to the vessel's diameter to avoid under-occlusion or tissue damage.

To address this, the pipeline incorporates the reference vessel diameter d_0, obtained from the pre-operative model at the clamping location $\mathbf{p}_c$. For the initial implementation, we assume that localized deformation does not significantly alter the diameter, such that

$$d(t) \approx d_0, \tag{5.4}$$

an assumption justified by the relative stiffness and small cross-sectional deformation of arteries under typical surgical conditions. Nevertheless, the framework is compatible with real-time estimation of $d(t)$ from intraoperative imaging for future use.

Using the tracked point $\mathbf{p}_c(t)$ and the estimated diameter $d(t)$, the system computes the required tool aperture $a(t)$ to safely occlude the vessel:

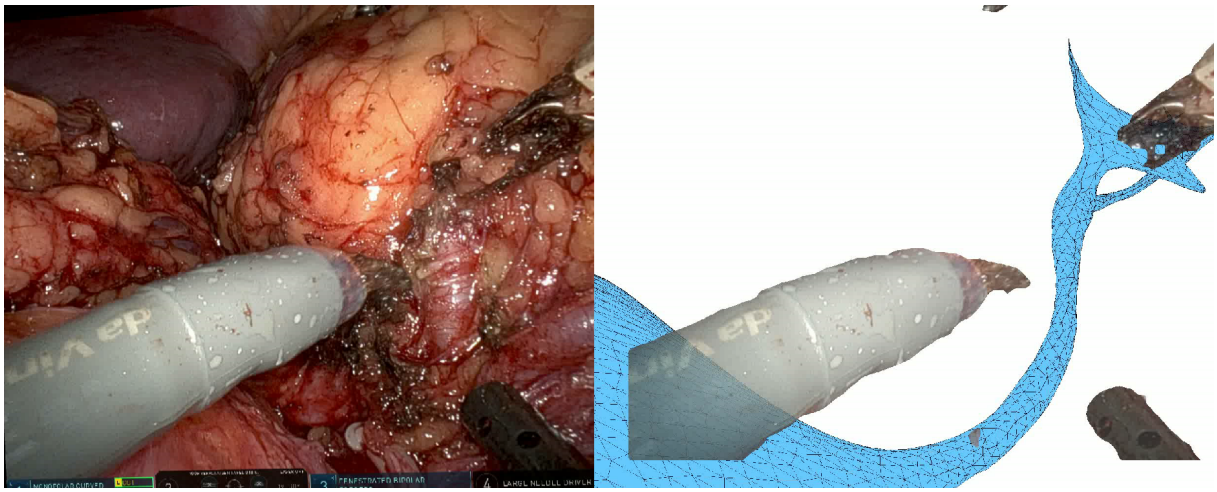
$$a(t) = f(d(t)), \tag{5.5}$$

where $f(\cdot)$ is a calibration function mapping a desired vessel diameter to the corresponding gripper opening. Thus, the robot receives two synchronized commands at each time step: the updated target position $\mathbf{p}_c(t)$ and the appropriate aperture $a(t)$ required for a safe and complete occlusion.

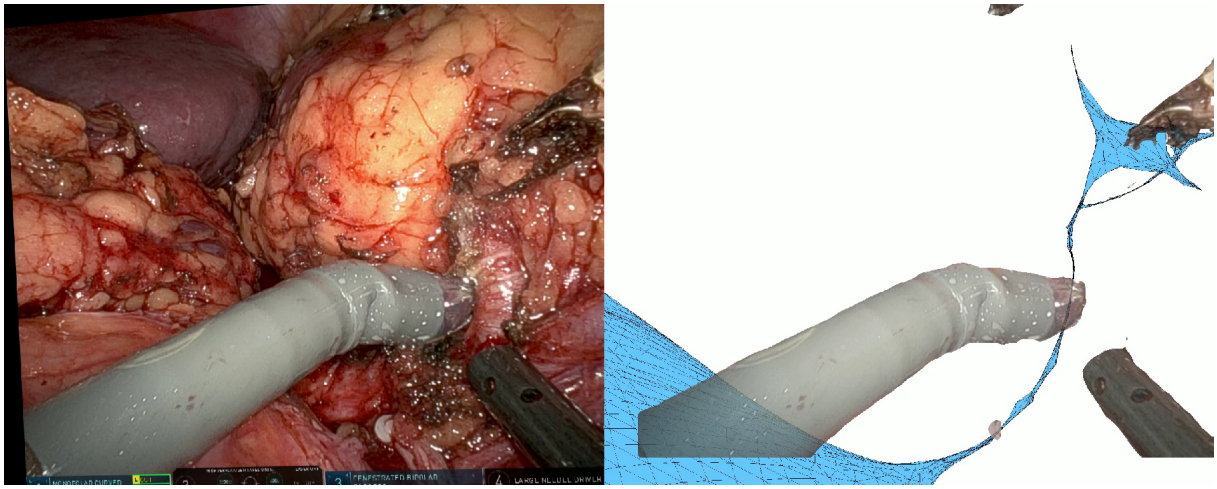
5.3.4 Conclusions and Next Steps

This implementation pipeline transforms continuous perceptual input into discrete, actionable robotic commands. By propagating the clamping point along the dynamically evolving artery centerline and adjusting the tool aperture based on the vessel diameter, the system achieves a deformation-aware clamping strategy grounded in anatomical realism. This integration provides robustness against tissue motion, ensures that the clamping action remains aligned with the intended target, and forms a crucial step toward autonomous surgical systems capable of operating safely in dynamic environments. At the current stage of development, this deformation-aware

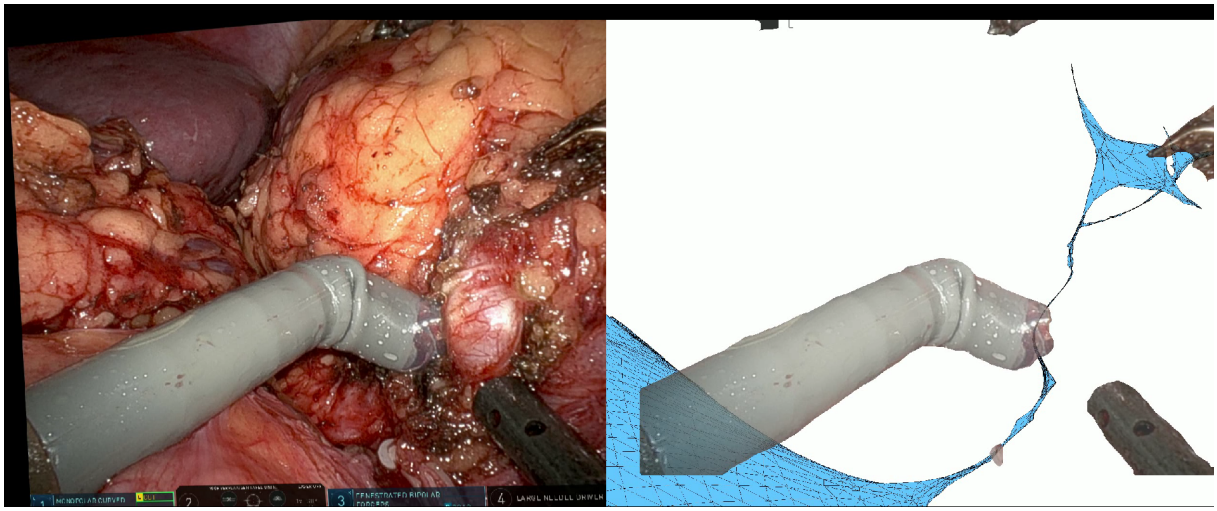
clamping framework has been validated only from a visual and geometric standpoint. Similar to the evaluation of deformation accuracy in the previous chapter, the propagated clamping points have been assessed on surgical videos to verify qualitative consistency with the observed arterial motion. However, no robotic execution, neither in virtual simulations nor on physical silicone phantoms, has yet been performed. Future work will involve deploying this pipeline on real and synthetic organ models to assess its geometric robustness, mechanical realism, and suitability for fully autonomous clamping tasks. Ultimately, deformation-aware clamping represents a crucial milestone toward fully autonomous surgical systems, where perception, reasoning, and actuation operate within a unified framework guided by real-time anatomical awareness.



((a)) Initial artery before deformation.



((b)) Tool approached artery branch (centerline) to move it.



((c)) Updated artery and clamping points after deformation.

Figure 5.1: Deformation-aware clamping points following arterial motion.

5.4 Deformation-Aware Tumor Cutting and Path Following

5.4.1 Introduction and motivations

The deformation tracking framework developed in the previous chapter demonstrated the feasibility of estimating non-rigid tissue motion using visual input alone. While simulation provides a controlled environment for testing algorithmic correctness, it cannot capture the full spectrum of physical phenomena encountered in real robotic interactions with soft tissue. Silicone phantoms introduce complexities such as sensor noise, lighting variability, surface reflections, non-linear elastic responses, local friction effects, and viscoelastic behavior that cannot be fully replicated in a virtual environment. Therefore, physical validation is a fundamental requirement for assessing the robustness, generalizability, and real-world applicability of the deformation-aware pipeline.

The core challenge addressed in this section is the preservation of a pre-planned surgical trajectory—specifically, the tumor cutting path—when the underlying organ undergoes dynamic and non-rigid deformation. As the robot interacts with the phantom or as external perturbations are applied, the tumor margin moves, rotates, and deforms. A robot executing a fixed trajectory will rapidly deviate from the anatomical target, compromising surgical accuracy. By integrating the learned deformation field $\mathbf{D}(\mathbf{x}, t)$ into the trajectory execution pipeline, the robot can continuously align its target pose with the evolving geometry of the phantom.

This experimental section constitutes the first physical validation of the integration between the imitation learning policy from Chapter 2, which provides the motion primitives and task execution framework, and the deformation tracking system from Chapter 4, which updates target poses in real time. Together, they enable adaptive surgical execution in deforming environments. Tumor cutting was chosen as the experimental task as it represents a complex, clinically relevant procedure involving multi-step tool–tissue interactions and strong dependence on accurate geometric tracking.

5.4.2 Experimental Setup and Path Definition

Robotic Hardware and Tooling All experiments were performed using a KUKA colaborative robot (or UR5, depending on the final setup), operating in Cartesian impedance mode to ensure compliant interaction with the phantom. The robot was equipped with the custom actuated surgical tool described in the previous section, enabling programmable cutting motions with highly repeatable jaw actuation. A second, passive auxiliary tool—a standard laparoscopic grasper mounted on a robotic or manually operated holder—was used to apply controlled perturbations to the silicone organ during the tests.

Visual Sensing The primary sensing modality was an Intel RealSense depth camera, selected for its compact form factor, integrated RGB-D fusion, and suitability as an analogue for laparoscopic monocular or stereo imaging. The camera was positioned to achieve a stable view of the operative field while minimizing occlusion.

Silicone Organ Phantoms The physical phantoms consisted of silicone kidneys with integrated tumors, manufactured from patient-derived CT scans. Tumor and kidney volumes were segmented, converted into high-resolution meshes, and used to fabricate silicone molds. Importantly, the silicone mixtures were selected to replicate physical properties—elasticity, hardness, and tactile response—that closely approximate those of real kidney tissue and tumor masses. Although silicone cannot perfectly reproduce all biomechanical characteristics of living tissue, the phantoms used in this work exhibit stiffness and deformation behavior sufficiently realistic for validating geometric deformation tracking and cutting performance.

Definition of the Cutting Path The cutting path was defined as the three-dimensional curve representing the intersection between the tumor and the surrounding kidney tissue, expanded by a 2 mm safety margin to emulate clinical practice. The path was discretized into uniform time steps and transformed into a series of waypoints. During execution, these waypoints were updated according to the deformation field:

$$\mathbf{p}(t) = \mathbf{p}_0 + \mathbf{D}(\mathbf{p}_0, t),$$

ensuring that the robot follows the anatomical boundary rather than a static trajectory.

5.4.3 Tool Orientation Policy Under RCM Constraints

Blade Orientation During Cutting

During tissue dissection, the robot must follow a predefined spatial trajectory while ensuring a physically and clinically valid cutting pose, all while respecting the Remote Center of Motion (RCM) imposed by the minimally invasive access port. We define a universal task-based orientation policy that determines the tool pose at each trajectory point by combining (i) the RCM constraint, (ii) the geometric structure of the cutting path, and (iii) a physically and clinically plausible slicing configuration.

Let $\mathbf{p}(s) \in \mathbb{R}^3$ denote the tool tip position at arc-length parameter s of the cutting trajectory, and let $\mathbf{r}_{\text{RCM}} \in \mathbb{R}^3$ denote the trocar entry point. The tool shaft axis is defined as the normalized vector

$$\hat{\mathbf{z}}(s) = \frac{\mathbf{p}(s) - \mathbf{r}_{\text{RCM}}}{\|\mathbf{p}(s) - \mathbf{r}_{\text{RCM}}\|}, \tag{5.6}$$

which ensures that the instrument pivots about the RCM without exerting lateral forces on the trocar. This construction fixes two rotational degrees of freedom; the remaining rotational freedom is a roll angle $\theta(s)$ around $\hat{\mathbf{z}}(s)$.

At each trajectory point, we define a local cutting frame using task geometry:

- **Feed direction (trajectory tangent):**

$$\hat{\mathbf{t}}(s) = \frac{d\mathbf{p}(s)/ds}{\|d\mathbf{p}(s)/ds\|}, \tag{5.7}$$

- **Surface normal (penetration direction):**

$$\hat{\mathbf{n}}(s), \tag{5.8}$$

  obtained from the anatomical surface model or preoperative imaging,

- **Cutting plane:** the plane spanned by $\hat{\mathbf{t}}(s)$ and $\hat{\mathbf{n}}(s)$.

The cutting plane captures the directions relevant for slicing and penetration into tissue.

Let $\hat{\mathbf{e}}(s)$ denote the direction of the blade edge. We constrain the blade to lie within the cutting plane and adopt a constant *slicing angle* $\alpha \in (0, \frac{\pi}{2})$ relative to the feed direction:

$$\hat{\mathbf{e}}(s) = \cos(\alpha)\,\hat{\mathbf{t}}(s) + \sin(\alpha)\,\hat{\mathbf{n}}_{\perp}(s), \tag{5.9}$$

where $\hat{\mathbf{n}}_{\perp}(s)$ is the normalized projection of $\hat{\mathbf{n}}(s)$ onto the cutting plane. The roll angle $\theta(s)$ around the shaft is then chosen so that the blade edge aligns with $\hat{\mathbf{e}}(s)$. This guarantees that advancing along the trajectory induces a non-zero component of motion *across* the blade edge, producing effective slicing rather than pure sliding.

To avoid configurations in which the lateral tool axis becomes perpendicular to the cutting path, undesirable under RCM kinematics and potentially unsafe in a surgical setting, we restrict the roll angle to a feasible interval:

$$\theta(s) \in [\theta_{\min}, \theta_{\max}], \tag{5.10}$$

chosen to avoid trocar loading, robot joint limits, and collisions.

Applicability to Soft and Deformable Tissue

Although the orientation policy described above is formulated in terms of path tangents and surface normals, it does not rely on rigid-body assumptions. The same geometric construction applies to soft and deformable organs, provided that the local geometric features used to compute the cutting frame are updated online. In minimally invasive surgery, tissue deformation occurs primarily in the vicinity of the tool tip; therefore, the location of the cutting point and the corresponding surface normal may change as the tool interacts with the organ.

Let $\tilde{\mathbf{p}}(s,t)$ denote the *deformed* position of the point on the cutting trajectory at time t, obtained from our intraoperative deformation pipeline. In our implementation, this is estimated from stereo depth, dense optical flow, or a learned deformation model, depending on the sensing modality available. Similarly, let $\hat{\mathbf{n}}(s,t)$ denote the instantaneous surface normal at the deformed location. The orientation policy directly incorporates these updated quantities by redefining the feed direction and surface normal as

$$\hat{\mathbf{t}}(s,t) = \frac{\partial \tilde{\mathbf{p}}(s,t)/\partial s}{\|\partial \tilde{\mathbf{p}}(s,t)/\partial s\|}, \tag{5.11}$$

$$\hat{\mathbf{n}}(s,t), \tag{5.12}$$

which together determine the instantaneous cutting plane. The tool shaft axis remains constrained by the RCM via

$$\hat{\mathbf{z}}(s,t) = \frac{\tilde{\mathbf{p}}(s,t) - \mathbf{r}_{\mathrm{RCM}}}{\|\tilde{\mathbf{p}}(s,t) - \mathbf{r}_{\mathrm{RCM}}\|}, \tag{5.13}$$

and the blade is oriented at the prescribed slicing angle α within the updated cutting plane. Thus, deformation affects only the inputs to the orientation policy, not its structure.

From a physical perspective, this formulation ensures that the blade responds to real-time local tissue geometry: as the surface bends or compresses under contact, the orientation is recomputed so that the slicing direction remains consistent with the true instantaneous interface between tumor and kidney. From a medical perspective, this avoids excessive shear or gouging and emulates the way surgeons naturally adjust their instrument orientation in response to tissue compliance. Mathematically, the policy remains valid because it depends solely on the instantaneous differential geometry of the deformed surface, and not on any assumption of rigidity.

Our deformation pipeline provides both the updated point location $\tilde{\mathbf{p}}(s,t)$ and the corresponding normal $\hat{\mathbf{n}}(s,t)$ in real time, enabling the robot to adjust its orientation continuously as tissue deforms. Therefore, the proposed method is directly applicable to soft-body interaction tasks: the trajectory is executed on the deformed organ rather than the preoperative model, while the RCM constraint and orientation structure remain unchanged. This ensures that the cutting action remains feasible, safe, and geometrically valid even in the presence of substantial, tool-induced soft tissue deformation.

Orientation Policy During Non-Cutting Motions

During non-cutting phases (e.g., returning from the end of one depth curve to the start of the next), the primary objective is safe traversal rather than maintaining a specific blade orientation.

The shaft axis continues to satisfy (5.6), ensuring RCM validity throughout the transition.

Let $d(\mathbf{p}(s))$ denote the minimum distance between the tool and the tissue surface, and let $d_{\mathrm{safe}} > 0$ be a predefined clearance. During non-cutting motion we require

$$d(\mathbf{p}(s)) \geq d_{\mathrm{safe}}, \tag{5.14}$$

ensuring that the blade does not accidentally contact or damage tissue. Roll adjustments $\theta(s)$ and small RCM-preserving repositioning motions may be used to maintain this clearance.

Because the complete trajectory is known preoperatively, the robot transitions smoothly toward the cutting orientation required at the start of the next incision. Let θ_{next} denote the roll angle needed at the next cutting point. We interpolate smoothly as

$$\theta(s) = (1 - \lambda(s))\,\theta_{\text{current}} + \lambda(s)\,\theta_{\text{next}}, \tag{5.15}$$

where $\lambda(s)$ increases monotonically from 0 to 1 along the non-cutting segment. This interpolation maintains continuity of the tool pose while ensuring that the correct slicing configuration is reached at the onset of the next cutting motion.

Geometric Feasibility

Since the shaft axis $\hat{\mathbf{z}}(s)$ is fixed by the RCM constraint, the only remaining rotational freedom is the roll angle around $\hat{\mathbf{z}}(s)$. A feasible orientation exists if and only if the desired blade-edge direction $\hat{\mathbf{e}}(s)$ is not parallel to the shaft axis:

$$\hat{\mathbf{e}}(s) \cdot \hat{\mathbf{z}}(s) \neq \pm 1. \tag{5.16}$$

The approach becomes geometrically infeasible only in the degenerate case where both the trajectory tangent and the surface normal align with the trocar axis,

$$\hat{\mathbf{t}}(s) \parallel \hat{\mathbf{z}}(s) \quad \text{and} \quad \hat{\mathbf{n}}(s) \parallel \hat{\mathbf{z}}(s), \tag{5.17}$$

causing the cutting plane to collapse into a line and forcing the desired blade orientation to lie along $\hat{\mathbf{z}}(s)$, which cannot be achieved by rotation about the shaft. This configuration is not encountered in typical minimally invasive procedures, as organ curvature and the oblique approach dictated by the RCM prevent the path tangent and surface normal from becoming colinear with the trocar direction.

5.4.4 Experiments

A three-stage protocol was designed to evaluate performance under progressively more realistic deformation scenarios.

Stage 1: Passive Deformation Tracking Validation The first stage assessed the accuracy and responsiveness of the deformation estimation pipeline independently of robot motion. External perturbations, including pushing, pulling, twisting, and lifting, were applied with the auxiliary tool. The deformation field was estimated continuously, and accuracy of the deformation was measured as in chapter 4:

Evaluation of deformation accuracy was performed using the same metrics described in the previous chapter, specifically:

- photometric consistency between the rendered and observed RGB appearance,
- silhouette consistency measuring alignment between predicted and observed object masks,
- depth consistency computed as the per-pixel deviation between predicted and measured depth,
- flow consistency evaluating the coherence between estimated optical flow and deformation-induced motion.

Together, these metrics capture the geometric and visual fidelity of the deformation field $\mathbf{D}(\mathbf{x}, t)$ on the physical silicone phantom.

Stage 2: Adaptive Path Following Under External Perturbation The second stage introduced dynamic adaptation without tool–tissue contact. The main tool followed the cutting path in free space while the auxiliary tool applied localized perturbations to the organ. The deformation-aware system updated the target pose in real time. Performance in this stage was assessed using three quantitative metrics:

- tracking error, defined as the Euclidean distance $\|p_{\text{des}}(t) - p_{\text{curr}}(t)\|$ in millimeters,
- deformation-to-tracking-error ratio, quantifying how well the robot compensates for the applied perturbation,
- recovery time, defined as the time required for the robot to realign with the deformed path after a perturbation.

A low tracking error and deformation/tracking ratio close to one indicate successful trajectory adaptation despite significant tissue motion.

Stage 3: Autonomous Tumor Cutting with Active Deformation The final stage evaluated full cutting execution. The actuated tool performed the cutting motion along the dynamically updated path, while the auxiliary tool retracted or stabilized tissue. Deformation was thus induced both by the cutting action and the auxiliary manipulation, creating a realistic testing environment for the integrated perception–reasoning– action loop. Performance in this stage was measured as in stage 2.

5.4.5 Discussion: Force Control and Justification

Precise force control is important for surgical tasks, but the current tool lacks an embedded force sensor. Thus, cutting was performed using position and velocity control:

- A constant feed rate was used for progressive cutting.
- Cutting depth was fixed relative to the local surface normal.

This choice is justified because the silicone phantom has homogeneous stiffness, reducing the need for adaptive force modulation. Additionally, the scientific focus of this work is on geometric and kinematic adaptation rather than force regulation. Fixed-depth cutting is also commonly used as a baseline in soft-tissue robotics research.

5.4.6 Conclusion and Future Work

The deformation tracking pipeline was evaluated through a set of experiments designed to quantify its accuracy, robustness, and suitability for downstream integration with adaptive surgical execution. These experiments assessed the performance of the system across controlled visual scenarios and increasingly realistic physical phantom settings, with the goal of verifying that the estimated deformation field $\mathbf{D}(\mathbf{x}, t)$ provides a reliable basis for dynamic trajectory updates and high-level surgical reasoning.

Accuracy was assessed using a combination of photometric consistency, silhouette alignment, depth agreement, and flow coherence. These metrics quantify how closely the predicted deformation matches the observed organ motion on a frame-to-frame basis and capture both local and global aspects of tissue displacement.

Across the evaluated sequences, the deformation estimates demonstrated low reconstruction error, smooth temporal evolution, and strong correspondence with visually observed motion. Photometric and depth-based metrics remained stable under varying illumination, while consistency across silhouette and flow fields confirmed the ability of the pipeline to track complex, non-uniform deformations of the organ surface.

Additional experiments examined the system's robustness under more dynamic conditions, including externally induced non-rigid perturbations. Local indentations and broader organ shifts were introduced to simulate the types of deformation expected during surgical manipulation. The predicted deformation fields successfully captured both global trajectories and subtle surface deviations, without exhibiting drift or instability over time.

These findings demonstrate the pipeline's ability to maintain reliable estimates even when the tissue undergoes rapid or spatially irregular motion, reinforcing its potential for integration with adaptive control strategies.

Visual inspection of the deformation maps confirmed that the estimated displacement fields were spatially coherent and consistent with the applied deformation. Global shifts of the organ were smoothly propagated across the entire surface, while localized deformations were accurately captured around regions of external contact. Overlays between predicted and observed shapes indicated close geometric agreement and demonstrated the method's capacity to model both rigid-body and elastic components of motion.

The system also showed resilience to realistic visual challenges such as shadows, minor occlusions, and texture inconsistencies. This robustness is essential for real-world surgical settings, where visual conditions are rarely ideal.

Beyond deformation accuracy, a key objective of this evaluation was to assess the readiness of the pipeline for use in downstream reasoning and execution modules. Preliminary experiments involving continuous point propagation demonstrated that the deformation field can reliably update anatomical landmarks, vascular centerline points, and cutting trajectories. The temporal stability and spatial fidelity of the deformation field indicate that it provides sufficiently accurate inputs for adaptive trajectory correction.

While full robotic execution tests are planned as the next development phase, the results obtained so far confirm that the perceptual framework meets the fidelity and consistency requirements needed for integration into deformation-aware surgical tasks. The strong performance across multiple visual and geometric criteria suggests that the system is well positioned for the subsequent execution-level experiments.

Although the current results are encouraging, several opportunities for refinement remain. The reliance on visual data alone limits the pipeline's ability to infer deeper mechanical properties such as tissue stiffness or internal tension, which may be important for tasks involving significant force application. Integrating complementary sensing modalities, particularly force measurements from the actuated tool, will enhance the model's ability to interpret tool–tissue interactions.

Furthermore, while the present experimental conditions include realistic phantom geometries and controlled perturbations, future evaluations will incorporate more complex task scenarios and richer interaction dynamics. This will allow a more comprehensive assessment of the system's performance under conditions closer to full surgical execution.

5.5 Conclusion: From Awareness to Automation

This chapter consolidates the methodological components developed throughout the thesis and demonstrates how learning, reasoning, and perception can be integrated into a coherent system for autonomous surgical execution. The central theme of the chapter is the progression from isolated algorithmic capabilities, in particular surgical reasoning through geometric and anatomical analysis, and deformation tracking through vision-based perception, to a unified framework capable of acting intelligently in realistic, deformable surgical environments. This integration represents a significant step toward achieving context-aware autonomy in robotic surgery.

The chapter presented two major applications in which deformation awareness plays a critical role. The first is deformation-aware tumor cutting, where the robot successfully maintained alignment with a dynamically deforming cutting path on silicone phantoms derived from pa-

tient CT scans. This task constitutes one of the most demanding forms of geometric adaptation, as the robot must continuously update its trajectory in response to externally applied perturbations or self-induced deformation caused by the cutting action. The second application is deformation-aware selective clamping. By propagating the pre-computed clamping points through the deformation field and updating arterial centerline geometry in real time, the system demonstrated how high-level surgical planning can be preserved even as the underlying anatomy shifts. These results show that the deformation pipeline enables the continuous maintenance of spatial consistency between surgical intent and real-world anatomy.

A crucial enabling element of these achievements is the development of a custom actuated surgical tool. Standard laparoscopic instruments lack motorized control and cannot be integrated directly with research robots. The tool developed in this work resolves this limitation by providing precise, programmable actuation suitable for both autonomous execution and high-quality data acquisition. It serves as the physical interface between the robotic system and the surgical environment, making it possible to test adaptive cutting and clamping strategies on physical phantoms. The tool therefore plays a foundational role in bridging algorithmic developments with real-world robotic interaction.

Beyond its specific experimental outcomes, the chapter offers broader technical and scientific impact. By performing controlled, vision-guided experiments on silicone organs manufactured from real patient anatomies, the work directly addresses the sim-to-real gap that hinders surgical robotics research. Many algorithms in surgical autonomy remain validated only in simulation or highly constrained environments. The experimental platform introduced here provides a reproducible, open-architecture environment in which learning-based, perception-based, and reasoning-based components can be tested on realistic tasks. This is particularly valuable given the substantial difficulty of collecting high-quality data in surgical settings, the constraints imposed by single-view endoscopic imaging, and the practical challenges of operating on deformable tissues. The ability to generate, track, and follow deformation on real phantoms is a major contribution to the research community.

Despite these advances, several limitations and challenges remain. The silicone phantoms used in the experiments, although manufactured from patient CT scans, still approximate mechanical properties and lack physiological phenomena such as perfusion, heterogeneous stiffness, or pulsation. Also the appearence of the environments we created with phantoms are still far from being close to the real surgical ones. Experimental setups involving lighting variation, tool occlusion, and narrow viewing angles present additional challenges for reliable perception. These constraints point to the inherent complexity of autonomous surgical robotics and reinforce the importance of the progress made so far.

Future work on selective clamping should focus on completing the full loop of deformation-aware autonomous execution on physical phantoms. This includes not only tracking the clamping points but also performing the actual grasping and clamp placement actions using the actuated end-effector. Incorporating the continuous estimation of arterial diameter and modeling the mechanical interaction between the tool and the vessel will be essential for precise and safe operation. The successful completion of these experiments would represent a significant step toward a fully automated clamping workflow.

In the domain of tumor cutting and more broadly in perception-based manipulation, the highest priority for future work is the integration of force and torque sensing. The current system uses a position- and velocity-controlled strategy, which is sufficient for homogeneous silicone but insufficient for real tissues, where cutting resistance varies spatially. A force-controlled autonomous cutting approach would allow the policy to adapt dynamically to tissue stiffness and detect unsafe interactions. Expanding the task repertoire to include procedures such as cauterization, blunt and sharp dissection, and tensioned manipulation will further challenge and extend the deformation pipeline. Improvements in deformation estimation—such as physics-informed modeling or multi-modal sensor fusion— may reduce drift and increase robustness.

Finally, transitioning from silicone phantoms to ex-vivo or in-vivo tissues is an essential long-term step for capturing realistic biophysical behavior and validating safety in clinically relevant conditions.

Ultimately, the work presented in this chapter demonstrates that perceptual awareness of deformation is not merely a component of surgical autonomy but a foundational requirement. Without accurate and timely understanding of how tissues move, stretch, and deform, neither learned policies nor high-level reasoning can be executed safely. The integration of learning, reasoning, and perception developed in this chapter forms the basis of a future surgical robotic system in which safety, adaptability, and intelligence are inherently linked. As the underlying algorithms, mechanisms, and sensing modalities evolve, the guiding principle will remain constant: autonomy must always be developed hand-in-hand with robust safeguards and an unwavering commitment to safe interaction with soft tissue.

Chapter 6

Conclusion

6.1 Thesis Synthesis and Foundational Achievement

The central hypothesis underlying this thesis is that safe, adaptive surgical autonomy can only be achieved by integrating four tightly connected capabilities: learning, reasoning, perceiving, and acting. Learning provides the robot with expert-derived skill and trajectory templates; perceiving allows it to understand and interpret the dynamic, deformable environment in which soft-tissue surgery unfolds; reasoning enables the system to use perceptual information to update high-level plans in real time; and acting translates those plans into precise, reliable physical motion through an appropriate surgical tool. The work presented in this thesis has advanced each of these pillars individually and, more importantly, demonstrated their integration into a unified framework capable of executing adaptive surgical subtasks on soft-tissue phantoms. This integration constitutes the foundational achievement of the thesis.

The research journey began with the acquisition of expert demonstrations and the development of an imitation learning framework capable of predicting complex surgical trajectories involving cutting and grasping. Building on this foundation, the thesis introduced a real-time soft tissue deformation tracking pipeline designed to estimate non-rigid organ motion from monocular or RGB-D video. The deformation field produced by this pipeline was then incorporated into a reasoning mechanism that continuously updates key surgical targets, such as clamping points along arterial centerlines and resection paths along tumor boundaries. This perceptual reasoning enables the robotic system to maintain alignment between its planned actions and the current anatomical state, even as the scene evolves under external and self-induced deformation.

Ultimately, the thesis demonstrates the first integrated system capable of perceptually adaptive robotic execution of clinically relevant soft-tissue tasks on physical phantoms. Through the combined use of imitation learning, deformation-aware perception, geometric reasoning, and precise physical actuation, the system performs both selective clamping and tumor cutting in the presence of complex non-rigid motion. This achievement establishes a new benchmark for autonomous soft-tissue surgical research and provides a structured methodology for future advances in safe, intelligent surgical robotics.

6.2 Summary and Contribution Mapping

The work developed in this thesis is organized according to four functional pillars: learning, reasoning, perceiving, and acting. Each chapter addresses one or more of these pillars, and the collective contributions form a comprehensive framework for adaptive surgical autonomy.

6.2.1 Chapter 2: Learning from Demonstrations

Chapter 2 presented the development of an imitation learning framework for learning expert surgical trajectories, focusing on tumor cutting and tissue manipulation. The work began with datasets collected in a virtual environment, where the flexibility of simulation allowed rapid generation of diverse scenarios while ensuring controlled conditions for early experimentation. The initial policy learned to predict tool positions, orientations, and binary cutting states from synchronized RGB-D frames, along with a temporal window of past observations and actions. This formulation enabled the model to capture temporal dependencies and reproduce multi-stage surgical behavior.

The second part of the chapter extended this framework to real-world physical phantoms, where demonstrations were recorded using a clinical-grade robot and RGB-D imaging system. This transition required addressing the sim-to-real gap, for which a domain adaptation strategy was incorporated into the learning pipeline. The resulting policy demonstrated the ability to generalize across variations in phantom geometry, tumor location, and lighting conditions.

In the LRPA paradigm, this chapter corresponds primarily to the Learning component, providing the foundational surgical skill that the subsequent modules adapt using perception and reasoning.

The contributions of this chapter include:

1. A domain-adaptive imitation learning pipeline for continuous surgical trajectories.
2. A demonstration dataset creation methodology for virtual and physical phantoms.
3. A trajectory prediction model capable of reproducing complex, multi-step surgical subtasks.
4. Real-world execution tests showing the feasibility of deploying learned policies on a robotic system.

6.2.2 Chapter 3: Surgical Reasoning for Selective Clamping

Chapter 3 developed the reasoning framework that enables the robot to update high-level surgical targets based on evolving anatomical conditions. The chapter focused on the selective clamping task in robot-assisted partial nephrectomy, where precise localization of arterial clamping points is essential for minimizing ischemic injury while preserving oncological margins.

Starting from pre-operative vascular models, the chapter introduced the geometric and anatomical reasoning required to identify safe clamping candidates along the renal arterial tree. The reasoning pipeline accounts for tumor location, perfusion territories, surgical accessibility and anatomical constraints. The chapter also formalized the representation of clamping points as a set of anatomically anchored targets that must be preserved even as the kidney undergoes intraoperative motion.

In the LRPA framework, Chapter 3 corresponds to Reasoning: it determines which targets must be tracked and updated during surgery and defines the logic connecting geometric anatomy to surgical intent. The methods developed here lay the conceptual foundation for later integration of real-time perception and robotic execution.

The contributions of this chapter include:

1. A complete reasoning framework for identifying and validating selective clamping points.
2. A formal definition of clamping targets anchored to vascular centerlines.
3. Integration of anatomical constraints and surgical accessibility into target selection.
4. Preparation of the reasoning layer for downstream integration with deformation-aware perception.

6.2.3 Chapter 4: Percieving - Soft Tissue Deformation Tracking

Chapter 4 presented the perceptual backbone of the thesis: a real-time soft tissue deformation tracking pipeline capable of estimating non-rigid motion from video. Using RGB-D data, the pipeline combined visual consistency metrics, depth alignment, silhouette matching and optical flow estimation to produce a deformation field for the entire organ surface. This deformation field captures both global kidney motion and local non-rigid deformation caused by tool interaction or external manipulation.

The pipeline was validated using silicone phantoms under controlled deformations, demonstrating accurate estimation of tissue displacement and robustness to visual noise. The use of CT-derived anatomies ensured that the phantoms closely replicate real patient geometries, although biological fidelity remains an open challenge.

Within the LRPA paradigm, this chapter corresponds to Perceiving: it provides the real-time situational awareness required for adaptive execution and enables the reasoning layer to update targets based on the current anatomical configuration.

The contributions of this chapter include:

1. A Deformation tracking pipeline based on RGB-D input.

2. A unified set of visual and geometric metrics for evaluating deformation accuracy.

3. Validation experiments on silicone phantoms manufactured from CT data.

4. Integration of the deformation field into subsequent reasoning and execution modules.

6.2.4 Chapter 5: Acting – Deformation-Aware Clamping, Tumor Cutting, and Surgical Tool Development

Chapter 5 brought together the learning, reasoning and perception components developed in the previous chapters and applied them to clinically relevant tasks: deformation-aware selective clamping and deformation-aware tumor cutting.

For selective clamping, the deformation field was used to update clamping points along the renal arterial centerline. The update rule $p(t) = p_0 + D(p_0, t)$ ensured that the target location for clamping remained aligned with the artery's true position despite geometric changes. Although autonomous robotic clamping was not executed, the perceptual reasoning and update logic were validated through video-based experiments.

For tumor cutting, the same update logic was used to deform the cutting trajectory corresponding to the tumor–kidney intersection curve. Path following error, deformation-to-tracking ratio and recovery time metrics demonstrated successful adaptation under dynamic conditions.

The contributions of this chapter include:

1. A deformation-aware reasoning framework for updating surgical targets during execution.

2. The first demonstration of adaptive tumor cutting on deformable silicone phantoms.

3. A detailed validation protocol for path following under dynamic deformation.

6.3 Limitations and Critical Analysis

Despite significant progress, the work presented in this thesis has several limitations that warrant further analysis. These limitations can be organized by module and by their broader implications for the LRPA paradigm.

6.3.1 Limitations of Imitation Learning

The imitation learning models developed in this work, while effective in controlled scenarios, face several challenges related to generalization, robustness, and data efficiency. Virtual demonstrations allowed large amounts of data to be collected, but discrepancies between simulated and real environments remain a major obstacle. Domain adaptation methods mitigated some of these discrepancies, yet the learned models continue to struggle when confronted with visual properties, lighting conditions, or geometric variations that fall outside the training distribution.

Real-world data scarcity remains a critical problem. Collecting demonstrations on physical phantoms is time-consuming, and while the actuated tool simplifies the process, the variety of trajectories remains limited. The models also lack explicit mechanisms for recovering from errors. Covariate shift can accumulate, causing trajectory drift or tool behavior inconsistent with expert demonstrations, and the system is not yet able to autonomously detect or correct these deviations.

6.3.2 Limitations of Deformation Tracking

The deformation tracking pipeline provides accurate estimates under most experimental conditions, but it relies primarily on visual cues and depth data. This sensory modality is inherently limited when dealing with occlusions, specular reflections or heavy tool–tissue contact. Without tactile or force-based sensing, the system cannot distinguish between visually similar deformations that may correspond to different underlying mechanical states. Furthermore, silicone phantoms approximate but do not fully capture the mechanical anisotropy, viscosity and nonlinearity of real biological tissues. As a result, the performance of the deformation model on real tissues remains uncertain.

The model also assumes that the deformation field is smooth and continuous, which is generally valid but may break down in cases of tearing, slipping, or sudden geometric changes. Latency in tracking, while low, can also introduce errors during rapid deformation.

6.3.3 Limitations of Deformation-Aware Clamping and Cutting

The main limitation of the clamping and cutting experiments is the absence of force control. Without a force or torque sensor integrated into the actuated tool, all actions are executed under position and velocity control, which is suitable for homogeneous silicone but inadequate for biological tissues. Cutting depth, contact force, and tool–tissue friction are important variables that the current system cannot regulate or measure. This restricts the complexity of cutting tasks that can be attempted safely.

Additionally, the experimental validation has so far been conducted exclusively on phantoms. While these phantoms provide realistic geometry, they do not simulate perfusion, tissue stiffness variability, bleeding or pulsation. These factors are critical to surgical safety and must be addressed before deploying such systems in real surgical contexts.

6.3.4 Limitations in the LRPA Paradigm

Beyond specific module limitations, the LRPA paradigm itself presents structural challenges. The learning-to-reasoning connection assumes that the high-level plan remains meaningful under shifting anatomy, yet errors in deformation tracking can cause incorrect target updates that propagate down the chain of execution. The reasoning-to-acting connection lacks explicit safety constraints, relying instead on the implicit safety encoded in the demonstration data. Finally, learning does not yet incorporate feedback from acting; the system does not learn from the outcomes of its adaptive behavior, meaning that improvements are not propagated back into the policy.

These limitations reflect the inherent difficulty of autonomous surgical robotics, where safety, adaptability and intelligence must coexist in a highly constrained domain.

6.4 Future Work and Research Roadmap

The limitations identified in this thesis offer clear directions for future advancement. The next stages of research should focus on improving individual modules, strengthening the connections between them, and extending the LRPA framework to more complex, realistic scenarios.

6.4.1 Future Work on Imitation Learning

Future work on the learning component should address the sim-to-real gap by incorporating domain randomization, generative models and physics-based simulation. These methods have the potential to produce policies that generalize better across diverse environments. New architectures such as large vision–language models or advanced transformer variants may further improve the model's ability to understand surgical scenes and predict actions.

Improving robustness is essential. Policies should be trained to detect when the executed behavior diverges significantly from expected trajectories, initiating recovery behaviors or safe stop conditions. This requires collecting failure examples and developing recovery strategies similar to those used in autonomous driving and locomotion.

Data collection must also evolve. More complex tasks such as dissection, suturing, or organ retraction should be incorporated into future datasets. Virtual environments should be enriched with physics-based modeling, realistic textures and adverse visual conditions to better emulate real surgery.

6.4.2 Future Work on Clamping and Deformation Applications

The selective clamping module should be extended to include full execution on physical phantoms. This requires refining the clamping point update logic, validating the deformation field under a wider variety of perturbations, and incorporating vessel diameter estimation using vision or tactile sensing.

Deformation tracking should incorporate force feedback from a sensor integrated into the actuated tool. Stiffness estimation and force-aware deformation inference would significantly improve accuracy and help adapt actions based on the mechanical response of tissues. Extending the pipeline to other organs and tasks, including cauterization, blunt dissection and suturing, will test its generality.

6.4.3 Future Work on the LRPA Paradigm

At the system level, future work should focus on closing the loop between action and learning, enabling adaptive execution to inform future policy updates. Explicit safety constraints should be integrated into the reasoning layer, including tissue strain monitoring and proximity constraints for critical structures. Finally, as a long-term goal, the entire framework should be validated on ex-vivo and eventually in-vivo tissues, providing evidence of robustness under realistic physiological conditions.

6.5 Final Remarks

This thesis has introduced a comprehensive, integrated methodology for adaptive autonomous execution in soft-tissue surgery, demonstrating its effectiveness on realistic physical phantoms. The combination of learning from expert demonstrations, perceiving soft tissue deformation, reasoning about evolving anatomy and acting through a custom surgical tool establishes a strong

foundation for future advances in surgical robotics. While significant challenges remain, the framework presented here offers a clear, structured path toward safe, intelligent and adaptive robotic systems that can assist surgeons in performing complex soft-tissue interventions with greater precision, consistency and safety.

Bibliography

[1] A. Attanasio, A. Fürst, P. Valdastri *et al.*, "Autonomy in surgical robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, pp. 651–679, 2021.

[2] A. Lee *et al.*, "Levels of autonomy in fda-cleared surgical robots: a systematic review," *npj Digital Medicine*, vol. 7, no. 1, p. 14, 2024.

[3] R. Khanna *et al.*, "The potential of autonomous surgery and challenges faced," *npj Digital Medicine*, 2025.

[4] A. Shademan, R. Decker, J. Opfermann *et al.*, "Supervised autonomous robotic soft tissue surgery," *Science Translational Medicine*, vol. 8, no. 337, p. 337ra64, 2016.

[5] J. Kim *et al.*, "Surgical robot transformer: Learning generalizable surgical policies from video and kinematic data," *Science Robotics*, 2025.

[6] T. Osa, J. Pajarinen, G. Neumann *et al.*, "An algorithmic perspective on imitation learning," *Foundations and Trends in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[7] T. Zhang *et al.*, "Deep imitation learning for complex manipulation tasks," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[8] NVIDIA, "What is imitation learning?" 2023. [Online]. Available: https://developer.nvidia.com/ai/imitational-learning

[9] A. Twinanda, S. Shehata, D. Mutter *et al.*, "Endonet: A deep architecture for surgical workflow recognition," *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 86–97, 2017.

[10] N. Jalal *et al.*, "A deep learning framework for surgical phase recognition using endoscopic video," *IFAC-PapersOnLine*, vol. 54, no. 15, pp. 210–217, 2021.

[11] A. C. P. Guédon *et al.*, "Deep learning for the recognition of surgical phases: A systematic review," *International Journal of Computer Assisted Radiology and Surgery*, 2021.

[12] A. Centeno López *et al.*, "Improving surgical phase recognition using self-supervised learning," *Scientific Reports*, vol. 15, p. 486, 2025.

[13] T. Carter, M. Sermesant *et al.*, "Application of soft tissue modelling to image-guided surgery," *Medical Image Analysis*, vol. 9, no. 6, pp. 587–604, 2005.

[14] D. Stoyanov, A. Darzi, and G.-Z. Yang, "Soft tissue deformation tracking for robotic assisted minimally invasive surgery," in *MICCAI*, 2005, pp. 309–316.

[15] R. Lastrico, F. Mastrogiovanni, P. Traverso, and D. A. Carfi, "Estimation of soft tissue deformation for robot-assisted surgery," 2021.

[16] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. Kim, "Supervised autonomous robotic soft tissue surgery," *Science Translational Medicine*, vol. 8, no. 337, p. 337ra64, 2016.

[17] E. Collaboration, A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. Wang, A. Kolobov, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. V. Frujeri, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Yang, G. Wang, H. Su, H.-S. Fang, H. Shi, H. Bao, H. B. Amor, H. I. Christensen, H. Furuta, H. Bharadhwaj, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Vakil, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K.-H. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, M. Z. Irshad, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Liu, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Tulsiani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Kumar, V. Vanhoucke, V. Guizilini, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, X. Liangwei, X. Li, Y. Pang, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Dou, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y.-H. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, Z. Fu, and Z. Lin, "Open x-embodiment: Robotic learning datasets and rt-x models," 2025. [Online]. Available: https://arxiv.org/abs/2310.08864

[18] N. Madapana, M. M. Rahman, N. Sanchez-Tamayo, M. V. Balakuntala, G. Gonzalez, J. P. Bindu, L. N. V. Venkatesh, X. Zhang, J. B. Noguera, T. Low, R. Voyles, Y. Xue, and J. Wachs, "Desk: A robotic activity dataset for dexterous surgical skills transfer to medical robots," 2019. [Online]. Available: https://arxiv.org/abs/1903.00959

[19] G. Furnari, C. Secchi, and F. Ferraguti, "Sequence-based imitation learning for surgical robot operations," *Artificial Intelligence Surgery*, vol. 5, pp. 103–15, 02 2025.

[20] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," *Foundations and Trends® in Robotics*, vol. 7, no. 1–2, p. 1–179, 2018. [Online]. Available: http://dx.doi.org/10.1561/2300000053

[21] M. Sahu, N. Pfreundschuh, T. Roß, S. Bodenstedt, and S. Speidel, "Sim-to-real domain adaptation using teacher-student learning for instrument segmentation," *Medical Image Analysis*, vol. 70, p. 102004, 2021.

[22] D. Zhang, Z. Wu, J. Chen, R. Zhu, A. Munawar, B. Xiao, Y. Guan, H. Su, W. Hong, Y. Guo, G. S. Fischer, B. Lo, and G.-Z. Yang, "Human-robot shared control for surgical robot based on context-aware sim-to-real adaptation," 2022. [Online]. Available: https://arxiv.org/abs/2204.11116

[23] T. Zhang, D. Zhao, H. Wang, L. Yang, and N. Simaan, "Context-aware human-robot shared control for surgical robot with sim-to-real adaptation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022, pp. 10 423–10 430.

[24] A. Pore, E. Tagliabue, M. Piccinelli, D. Dall'Alba, A. Casals, and P. Fiorini, "Learning from demonstrations for autonomous soft-tissue retraction," 2021. [Online]. Available: https://arxiv.org/abs/2110.00336

[25] F. Richter, S. Shen, F. Liu, J. Huang, K. Funk, Emily K. Orosco, Ryan and C. Yip, Michael "Autonomous robotic suction to clear the surgical field for hemostasis using image-based blood flow detection," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1383–1390, 2021.

[26] H. Saeidi, J. D. Opfermann, M. Kam, S. Wei, S. Léonard, M. H. Hsieh, J. U. Kang, and A. Krieger, "Autonomous robotic laparoscopic surgery for intestinal anastomosis," *Science robotics*, vol. 7, no. 62, p. eabj2908, 2022.

[27] B. Su, S. Yu, X. Li, Y. Gong, H. Li, Z. Ren, Y. Xia, H. Wang, Y. Zhang, W. Yao *et al.*, "Autonomous robot for removing superficial traumatic blood," *IEEE Journal of Translational Engineering in Health and Medicine*, pp. 1–9, 2021.

[28] H. Saeidi, N.-D. Le, Hanh D. Opfermann, Justin S. Léonard, A. Kim, H. Hsieh, Michael U. Kang, Jin and A. Krieger, "Autonomous laparoscopic robotic suturing with a novel actuated suturing tool and 3d endoscope," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 1541–1547.

[29] M. Ginesi, D. Meli, A. Roberti, N. Sansonetto, and P. Fiorini, "Autonomous task planning and situation awareness in robotic surgery," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 3144–3150.

[30] J. Han, J. Davids, H. Ashrafian, A. Darzi, S. Elson, Daniel and M. Sodergren, "A systematic review of robotic surgery: From supervised paradigms to fully autonomous robotic approaches," *The International Journal of Medical Robotics and Computer Assisted Surgery*, 2022.

[31] V. S. Bawa, G. Singh, F. KapingA, I. Skarga-Bandurova, E. Oleari, A. Leporini, C. Landolfo, P. Zhao, X. Xi, G. Luo, K. Wang, L. Li, B. Wang, S. Zhao, L. Li, A. Stabile, F. Setti, R. Muradore, and F. Cuzzolin, "The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods," *arXiv*, 2021.

[32] V. S. Bawa, G. Singh, F. KapingA, I. Skarga-Bandurova, E. Oleari, A. Leporini, C. Landolfo, A. Stabile, F. Setti, R. Muradore, and F. Cuzzolin, "Esad: Endoscopic surgeon action detection dataset," *arXiv*, 2020.

[33] J. Yoon, S. Hong, S. Hong, J. Lee, S. Shin, B. Park, N. Sung, H. Yu, S. Kim, S. Park *et al.*, "Surgical scene segmentation using semantic image synthesis with a virtual surgery environment," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2022, pp. 551–561.

[34] M. Carstens, F. M. Rinner, S. Bodenstedt *et al.*, "The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science," *Scientific Data*, vol. 10, p. 3, 2023.

[35] E. Colleoni, P. Edwards, and D. Stoyanov, "Synthetic and real inputs for tool segmentation in robotic surgery," *arXiv*, 2020.

[36] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *MICCAI Workshop: M2CAI*, 2014, p. 3, jIGSAWS dataset. [Online]. Available: https://cirl.lcsr.jhu.edu/research/hmm/datasets/jigsaws_release/

[37] I. Rivas-Blanco, C. J. Pérez-del Pulgar, A. Mariani, C. Quaglia, G. Tortora, A. Menciassi *et al.*, "A surgical dataset from the da vinci research kit for task automation and recognition," *arXiv*, 2021.

[38] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.

[39] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani *et al.*, "Transporter networks: Rearranging the visual world for robotic manipulation," in *Conference on Robot Learning (CoRL)*, 2021, pp. 726–747.

[40] S. Yang, W. Zhang, W. Lu, H. Wang, and Y. Li, "Cross-context visual imitation learning from demonstrations," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 5467–5473.

[41] H. M. Eraqi, M. N. Moustafa, and J. Honer, "End-to-end deep learning for steering autonomous vehicles considering temporal dependencies," *arXiv*, 2017.

[42] A. K. Tanwani, P. Sermanet, A. Yan, R. Anand, M. Phielipp, and K. Goldberg, "Motion2vec: Semi-supervised representation learning from surgical videos," *arXiv*, 2020.

[43] A. Murali, S. Sen, B. Kehoe, A. Garg, S. McFarland, S. Patil, W. D. Boyd, S. Lim, P. Abbeel, and K. 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)*, 2015, pp. 1202–1209.

[44] Z. Zhao, Tony V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv*, 2023.

[45] J. W. Kim, S. Schmidgall, A. Krieger, and M. Kobilarov, "Learning a library of surgical manipulation skills for robotic surgery," in *Bridging the Gap between Cognitive Science and Robot Learning in the Real World: Progresses and New Directions*, 2024. [Online]. Available: https://openreview.net/forum?id=fYRlaylCI3

[46] A. Karimi and A. Shojaei, "Measurement of the mechanical properties of the human kidney," *IRBM*, vol. 38, no. 5, pp. 292–297, 2017.

[47] A. Levillain, C. B. Confavreux, M. Decaussin-Petrucci, E. Durieux, P. Paparel, K. L.-B. Carval, L. Maillard, F. Bermond, D. Mitton, and H. Follet, "Mechanical properties of breast, kidney, and thyroid tumours measured by afm: Relationship with tissue structure," *Materialia*, 2022.

[48] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[49] J. Kim, C. Chi, Y. Lin, K. Xu, G. Hager, and R. Taylor, "Surgical robot transformer: Imitation learning for surgical tasks," in *Robotics: Science and Systems (RSS)*, 2024.

[50] E. Jang, A. Irpan, M. Khansari, D. Kappler, W. Zhou, M. Sharma, M. Kalakrishnan, and S. Levine, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning (CoRL)*, 2022, pp. 991–1002.

[51] K. O'Neill *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," *International Conference on Robotics and Automation (ICRA)*, 2024.

[52] G. Rosiello, P. Piazza, S. Puliatti, E. Mazzone, M. Amato, V. Tames, R. Farinha, R. D. Groote, C. Berquin, D. Develtere, C. Sinatti, A. Larcher, U. Capitanio, F. D'Hondt, P. Schatteman, A. Briganti, F. Montorsi, G. D. Naeyer, and A. Mottrie, "Simplified PADUA renal (SPARE) nephrometry score validation and long-term outcomes after robot-assisted partial nephrectomy," *Urologic Oncology*, vol. 40, no. 2, 2022.

[53] L. Xia, X. Wang, T. Xu, and T. J. Guzzo, "Systematic review and meta-analysis of comparative studies reporting perioperative outcomes of robot-assisted partial nephrectomy versus open partial nephrectomy," *Journal of Endourology*, vol. 31, no. 9, pp. 893–909, 2017.

[54] R. Schiavina, L. Bianchi, F. Chessa, U. Barbaresi, L. Cercenelli, S. Lodi, C. Gaudiano, B. Bortolani, A. Angiolini, F. M. Bianchi, A. Ercolino, C. Casablanca, E. Molinaroli, A. Porreca, R. Golfieri, S. Diciotti, E. Marcelli, and E. Brunocilla, "Augmented reality to guide selective clamping and tumor dissection during robot-assisted partial nephrectomy: A preliminary experience," *Clinical Genitourinary Cancer*, vol. 19, no. 3, pp. e149–e155, 2021.

[55] F. Greco, R. Autorino, V. Altieri, S. Campbell, V. Ficarra, I. Gill, A. Kutikov, A. Mottrie, V. Mirone, and H. van Poppel, "Ischemia techniques in nephron-sparing surgery: A systematic review and meta-analysis of surgical, oncological, and functional outcomes," *European Urology*, vol. 75, no. 3, 2019.

[56] R. Schiavina, L. Bianchi, M. Borghesi, F. Chessa, L. Cercenelli, E. Marcelli, and E. Brunocilla, "Three-dimensional digital reconstruction of renal model to guide preoperative planning of robot-assisted partial nephrectomy," *International Journal of Urology*, vol. 26, no. 9, pp. 931–932, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/iju.14038

[57] F. Porpiglia, D. Amparore, E. Checcucci, R. Autorino, M. Manfredi, G. Iannizzi, C. Fiori, E. R. Group *et al.*, "Current use of three-dimensional model technology in urology: a road map for personalised surgical planning," *European urology focus*, vol. 4, no. 5, pp. 652–656, 2018.

[58] F. Porpiglia, C. Fiori, E. Checcucci, D. Amparore, and R. Bertolo, "Hyperaccuracy three-dimensional reconstruction is able to maximize the efficacy of selective clamping during robot-assisted partial nephrectomy for complex renal masses," *European Urology*, vol. 74, no. 5, pp. 651–660, 2018.

[59] A. Leporini, E. Oleari, C. Landolfo, A. Sanna, A. Larcher, G. Gandaglia, N. Fossati, F. Muttin, U. Capitanio, F. Montorsi, A. Salonia, M. Minelli, F. Ferraguti, C. Secchi, S. Farsoni, A. Sozzi, M. Bonfè, N. Sayols, A. Hernansanz, A. Casals, S. Hertle, F. Cuzzolin, A. Dennison, A. Melzer, G. Kronreif, S. Siracusano, F. Falezza, F. Setti, and R. Muradore, "Technical and functional validation of a teleoperated multirobots platform for minimally invasive surgery," *IEEE Transactions on Medical Robotics and Bionics*, vol. 2, no. 2, pp. 148–156, 2020.

[60] G. De Rossi, M. Minelli, S. Roin, F. Falezza, A. Sozzi, F. Ferraguti, F. Setti, M. Bonfè, C. Secchi, and R. 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*, vol. 3, no. 3, pp. 714–724, 2021.

[61] F. Ferraguti, M. Minelli, S. Farsoni, S. Bazzani, M. Bonfè, A. Vandanjon, S. Puliatti, G. Bianchi, and C. Secchi, "Augmented reality and robotic-assistance for percutaneous nephrolithotomy," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4556–4563, 2020.

[62] B. K. J. Viertel, A. Naber, S. Hoffmann, D. Berwanger, L. Kessler, R. Khoramnia, and W. Nahm, "Automated vessel centerline extraction and diameter measurement in oct angiography," *Current Directions in Biomedical Engineering*, vol. 7, no. 2, pp. 195–198, 2021.

[63] K. Jarvis, S. Schnell, A. Barker, J. Garcia, R. Lorenz, M. Rose, V. Chowdhary, J. Carr, J. Robinson, C. Rigsby, and M. Markl, "Evaluation of blood flow distribution asymmetry and vascular geometry in patients with fontan circulation using 4-d flow mri," *Pediatric Radiology*, vol. 46, 2016.

[64] L. Antiga, B. Ene-Iordache, and A. Remuzzi, "Centerline computation and geometric analysis of branching tubular surfaces with application to blood vessel modeling," in *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 2003.

[65] L. Antiga and D. Steinman, "Robust and objective decomposition and mapping of bifurcating vessels," *IEEE Transactions on Medical Imaging*, vol. 23, no. 6, pp. 704–713, 2004.

[66] A. Bundy and L. Wallen, *Breadth-First Search.* Berlin, Heidelberg: Springer Berlin Heidelberg, 1984, pp. 13–13.

[67] F. Porpiglia, E. Checcucci, D. Amparore, F. Piramide, G. Volpi, S. Granato, P. Verri, M. Manfredi, A. Bellin, P. Piazzolla, R. Autorino, I. Morra, C. Fiori, and A. Mottrie, "Three-dimensional augmented reality robot-assisted partial nephrectomy in case of complex tumours (padua $\geq$ 10): A new intraoperative tool overcoming the ultrasound guidance," *European Urology*, vol. 78, no. 2, pp. 229–238, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0302283819308978

[68] S. Ali, Y. Espinel, Y. Jin, P. Liu, B. Güttner, X. Zhang, L. Zhang, T. Dowrick, M. J. Clarkson, S. Xiao, Y. Wu, Y. Yang, L. Zhu, D. Sun, L. Li, M. Pfeiffer, S. Farid, L. Maier-Hein, E. Buc, and A. Bartoli, "An objective comparison of methods for augmented reality in laparoscopic liver resection by preoperative-to-intraoperative image fusion from the miccai2022 challenge," *Medical Image Analysis*, vol. 99, p. 103371, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841524002962

[69] B. Koo, E. Özgür, B. Le Roy, E. Buc, and A. Bartoli, "Deformable registration of a preoperative 3d liver volume to a laparoscopy image using contour and shading cues," in *International conference on medical image computing and computer-assisted intervention.* Springer, 2017, pp. 326–334.

[70] Y. Espinel, L. Calvet, K. Botros, E. Buc, C. Tilmant, and A. Bartoli, "Using multiple images and contours for deformable 3D–2D registration of a preoperative CT in laparoscopic liver surgery," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 12, pp. 2211–2219, 2022. [Online]. Available: https://doi.org/10.1007/s11548-022-02774-1

[71] M. Labrunie, M. Ribeiro, F. Mourthadhoi, C. Tilmant, B. Le Roy, E. Buc, and A. Bartoli, "Automatic preoperative 3d model registration in laparoscopic liver resection," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 8, pp. 1429–1436, 2022.

[72] P. Guan, H. Luo, J. Guo, Y. Zhang, and F. Jia, "Intraoperative laparoscopic liver surface registration with preoperative ct using mixing features and overlapping region masks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 18, no. 8, pp. 1521–1531, 2023.

[73] J. S. Heiselman, J. A. Collins, L. W. Clements, J. A. Weis, A. L. Simpson, S. K. Geevarghese, T. P. Kingham, W. R. Jarnagin, and M. I. Miga, "Nonrigid registration for laparoscopic liver surgery using sparse intraoperative data," in *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 10576. SPIE, 2018, pp. 73–79.

[74] Z. Cao, L. Xie, and Y. Yang, "Non-rigid image-volume registration for human livers in laparoscopic surgery," *Quantitative Imaging in Medicine and Surgery*, vol. 15, no. 9, p. 8440, 2025.

[75] J. Zhou, B. Gao, K. Wang, J. Pei, P.-A. Heng, and J. Qin, "Landmark-free preoperative-to-intraoperative registration in laparoscopic liver resection," *IEEE Transactions on Medical Imaging*, 2025.

[76] M. Pfeiffer, C. Riediger, S. Leger, J.-P. Kühn, D. Seppelt, R.-T. Hoffmann, J. Weitz, and S. Speidel, "Non-rigid volume to surface registration using a data-driven biomechanical model," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 724–734.

[77] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissanayake, "Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 155–162, 2017.

[78] ——, "Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4068–4075, 2018.

[79] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, "Slam-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality," *Computer methods and programs in biomedicine*, vol. 158, pp. 135–146, 2018.

[80] N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel, "Live tracking and dense reconstruction for handheld monocular endoscopy," *IEEE transactions on medical imaging*, vol. 38, no. 1, pp. 79–89, 2018.

[81] X. Tang, H. Tao, Y. Qian, J. Yang, Z. Feng, and Q. Wang, "Real-time deformable slam with geometrically adapted template for dynamic monocular laparoscopic scenes," *International Journal of Computer Assisted Radiology and Surgery*, vol. 19, no. 7, pp. 1375–1383, 2024.

[82] L. Zhu, Z. Wang, J. Cui, Z. Jin, G. Lin, and L. Yu, "Endogs: Deformable endoscopic tissues reconstruction with gaussian splatting," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 135–145.

[83] S. Giannarou, M. Ye, G. Gras, K. Leibrandt, H. J. Marcus, and G.-Z. Yang, "Vision-based deformation recovery for intraoperative force estimation of tool–tissue interaction for neurosurgery," *International journal of computer assisted radiology and surgery*, vol. 11, no. 6, pp. 929–936, 2016.

[84] W.-J. Jung, K.-S. Kwak, and S.-C. Lim, "Vision-based suture tensile force estimation in robotic surgery," *Sensors*, vol. 21, no. 1, p. 110, 2020.

[85] M. Neidhardt, R. Mieling, M. Bengs, and A. Schlaefer, "Optical force estimation for interactions between tool and soft tissues," *Scientific reports*, vol. 13, no. 1, p. 506, 2023.

[86] K. Masui, N. Kume, M. Nakao, T. Magaribuchi, A. Hamada, T. Kobayashi, and A. Sawada, "Vision-based estimation of manipulation forces by deep learning of laparoscopic surgical images obtained in a porcine excised kidney experiment," *Scientific Reports*, vol. 14, no. 1, p. 9686, 2024.

[87] M. D. I. Reyzabal, D. Malas, S. Wang, S. Ourselin, and H. Liu, "Surgemod: Translating image-space tissue motions into vision-based surgical forces," *arXiv preprint arXiv:2406.17707*, 2024.

[88] Y. Wang, Z. Ye, M. Wen, H. Liang, and X. Zhang, "Transvfs: A spatio-temporal local–global transformer for vision-based force sensing during ultrasound-guided prostate biopsy," *Medical image analysis*, vol. 94, p. 103130, 2024.

[89] J. Zhan, J. Cartucho, and S. Giannarou, "Autonomous tissue scanning under free-form motion for intraoperative tissue characterisation," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 11 147–11 154.

[90] R. Lagneau, A. Krupa, and M. Marchal, "Active deformation through visual servoing of soft objects," in *2020 ieee international conference on robotics and automation (icra)*. IEEE, 2020, pp. 8978–8984.

[91] A. Krupa, M. de Mathelin, C. Doignon, J. Gangloff, G. Morel, L. Soler, and J. Marescaux, "Development of semi-autonomous control modes in laparoscopic surgery using automatic visual servoing," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2001, pp. 1306–1307.

[92] D. Arikan, P. Zhang, M. Sommersperger, S. Dehghani, M. Esfandiari, R. H. Taylor, M. A. Nasseri, P. Gehlbach, N. Navab, and I. Iordachita, "Real-time deformation-aware control for autonomous robotic subretinal injection under ioct guidance," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 10 531–10 537.

[93] N. Golse, A. Petit, M. Lewin, E. Vibert, and S. Cotin, "Augmented reality during open liver surgery using a markerless non-rigid registration system," *Journal of Gastrointestinal Surgery*, vol. 25, pp. 662–671, 2021.

[94] Z. Jiang, S. E. Salcudean, and N. Navab, "Robotic ultrasound imaging: State-of-the-art and future perspectives," *Medical Image Analysis*, vol. 89, p. 102878, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136184152300138X