CONACYT

UK Research and Innovation

UNIVERSITY OF LEEDS

Data-driven Spatially and Temporally Consistent Scene Representations for Robotic Minimally Invasive Procedures



Gerardo Emmanuel Loza Galindo

School of Computer Science

The University of Leeds

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

February 2026

Declaration of Authorship

I confirm that the work submitted in this thesis is my own, except where work forming part of jointly authored publications has been included. In such cases, my individual contribution and those of the co-authors are explicitly stated below. Appropriate acknowledgement has been given within the thesis where reference has been made to the work of others. I also acknowledge the use of ChatGPT-5 (OpenAI, https://chatgpt.com/) for proofreading and minor language refinement of the final draft. All scientific content, analysis, and conclusions remain my own.

- Chapter 3 is based on **Loza, G.**, Valdastri, P., Ali, S. (2024). Real-time surgical tool detection with multi-scale positional encoding and contrastive learning. Healthc. Technol. Lett. 11, 48–58. **(published)**
  Contributions:
  Loza, G.: Conceptualisation, methodology, software development, data processing, formal analysis, investigation, validation, visualisation, writing – original draft, review & editing.
  Ali, S. & **Valdastri, P.**: Supervision, methodology input, investigation support, writing – review & editing, resources, project administration, funding acquisition.

- Chapter 4 is based on **Loza, G.**, Hu, J., Jones, D., Ali, S., Valdastri, P. (2025). Surg-InvNeRF: Invertible NeRF for 3D tracking and reconstruction in surgical vision. https://www.arxiv.org/pdf/2508.09681. **(under review)**
  Contributions:
  Loza, G.: Conceptualisation, methodology, software development, data processing, formal analysis, investigation, validation, visualisation, writing – original draft, review & editing.
  Hu, J.: Investigation, writing – review & editing.
  Jones, D., **Ali, S.** & **Valdastri, P.**: Supervision, methodology input, investigation support, writing – review & editing, resources, project administration, funding acquisition.

The above publications are primarily my work. I led the conceptual development, implementation, experimentation, and analysis. Co-authors provided advisory input, supervision, and manuscript feedback.

- Chapter 5 is based on **Loza, G.**, Magro, M., Calme, B., Hu, J., Jones, D., Ruffaldi, E., De Momi, E., Ali, S., Valdastri, P. (2025). DRENDS: A Dataset for Depth in Robotic Endoscopy with Dynamic Scenarios. **(under review)**
  Contributions:
  Loza, G. & **Magro, M.**: Conceptualisation, protocol design, data collection, data processing, software development, validation, formal analysis, visualisation, writing – original draft, writing – review & editing.
  Calme, B.: Protocol design, tissue preparation, data collection, writing – review & editing.
  Hu, J.: Protocol design, data collection.
  Ruffaldi, E., **De Momi, E.**,**Jones, D.**, **Ali, S.** & **Valdastri, P.**: Supervision, methodology input, writing – review & editing, resources, project administration, funding acquisition.

The above publication represents a collaborative effort led by me. The specific contributions of each co-author are detailed above.

This copy has been supplied on the understanding that it is copyright material and that no quotation from my thesis may be published without proper acknowledgement. The right of Gerardo Emmanuel Loza Galindo to be identified as Author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

Acknowledgements

This thesis is the culmination of the support of many educators, funders, colleagues, friends, and family, without whom it would not have been possible. First, I would like to express my deepest gratitude to Professor Pietro Valdastri for the opportunity and support to pursue this PhD and for facilitating my integration into the STORMLAB community. Equally, I want to express my sincere gratitude to Dr Sharib Ali for his guidance, intellectual rigour, and continuous support throughout this research. His mentorship has been fundamental to the development of this thesis. I am also deeply thankful to Dr Dominic Jones for his invaluable technical insight, constructive feedback, and constant availability for discussion. His ability to challenge ideas while encouraging independent thinking has profoundly shaped this work.

I gratefully acknowledge the financial support provided by both British and Mexican institutions that made this research possible. I am deeply grateful for the public funding that supported this work, and for the trust placed in research institutions to advance knowledge for the benefit of society.

I want to thank all my colleagues in STORMLAB. Their company, ideas, conversations, jokes, and laughs made the PhD a genuinely enjoyable and enriching experience. I am also grateful to the lab technicians, Sam, Alfred, and Robyn, along with the administrative staff at the University of Leeds, for their invaluable support.

As part of the Leeds Medical AI CDT community, I would like to thank Professor David Hogg, Professor Vania Dimitrova, and Dr Owen Johnson for their tireless efforts in guiding and caring for all the students in this great community. Special thanks to Lewis and Ben for their friendship and support. Thanks to the CDT management team, particularly Hien, Sunjeev, and Azema.

Special thanks to Benjamin, Junlei, Nils, Vittorio, Nikita, Pedro, and Mattia, with whom I worked closely and from whom I learned valuable insights.

To my friends and family in the UK and Mexico, to Giuseppina, Jesús, Ahmad, Anna, Ian, Natacha, Lewis, Michael, Nikita, Paula, Erick, Donet, Jasiel, Eduardo, Vianney, Leonardo, Salvador, Rair, Humberto, Iram, and Mary. Thank you for your love and support. Despite the distance, I have always felt you close.

Mamá, gracias por tu apoyo incondicional y por todo lo que has sacrificado por mí. Tío Carlos, tía Isa, gracias por ser mi modelo a seguir y por estar siempre ahí para escucharme. Dennise, Diana, Alem, Sara, Hannia, Mariana, Emiliano, son mi motivación. Gracias por su cariño a pesar de estar tan lejos. Estoy seguro de que van a lograr grandes cosas.

Abstract

Robotic Minimally Invasive Procedures (MIP) generate large amounts of intraoperative data, including endoscopic video and instrument motion. Despite technological advances, current systems lack the ability to build a consistent and structured understanding of the surgical scene. This thesis addresses this limitation by investigating how visual information in robotic MIP can be organised into scene representations that remain coherent across space and time. The central premise is that surgical understanding depends not only on isolated task performance, but on how different visual representations (from object localisation to three-dimensional reconstruction) interact and support one another.

The thesis first introduces a representation-centric framework that formalises visual surgical perception as a multi-level process, clarifying how detection, tracking, motion estimation, and reconstruction are connected. A transformer-based detection model is then proposed to obtain robust object representations under realistic perturbations. Moving on, render-based implicit scene representations are developed to jointly model appearance, motion, and three-dimensional structure in deformable environments, enabling temporally consistent tracking and separation of camera motion from tissue deformation. Finally, the DRENDS dataset is presented to support quantitative evaluation of temporally consistent depth estimation in dynamic surgical scenarios.

Overall, this work demonstrates that effective surgical perception depends on how visual information is structured and integrated, and it contributes foundational methods and evaluation tools for Surgical Data Science.

Contents

Declaration of Authorship i

Acknowledgements iii

Abstract iv

1 Introduction 1
- 1.1 Minimally Invasive Procedures (MIP) . . . 1
- 1.2 Robotic MIP . . . 3
  - 1.2.1 Computer-Assisted Interventions . . . 4
  - 1.2.2 Robotic platforms and levels of autonomy . . . 5
- 1.3 Surgical Data Science . . . 9
  - 1.3.1 Visual Surgical Perception . . . 10
  - 1.3.2 Learning-Based Approaches for Visual Perception . . . 12
- 1.4 Visual Surgical Perception Limitations . . . 13
- 1.5 Aim of This Thesis . . . 13
  - 1.5.1 Objectives . . . 14
- 1.6 Contributions . . . 15
  - 1.6.1 Publications Arising from This Thesis . . . 15
  - 1.6.2 Relevant Collaborations . . . 15
- 1.7 Thesis Structure . . . 17

2 Foundations and Related Work 18
- 2.1 Local Primitives . . . 20
- 2.2 Identification of Surgical Objects . . . 21
  - 2.2.1 Challenges . . . 22
  - 2.2.2 Related Work . . . 23
- 2.3 Temporal Consistency . . . 27
  - 2.3.1 Challenges . . . 29
  - 2.3.2 Related Work . . . 30

2.4 Metric and Spatial Grounding . . . 34
2.4.1 Challenges . . . 35
2.4.2 Related Work . . . 36

3 Surgical Tool Representations 40
3.1 Problem definition . . . 41
3.2 Method . . . 42
3.2.1 Architecture details . . . 43
3.2.2 Loss functions . . . 44
3.3 Experiments and results . . . 47
3.3.1 Dataset . . . 47
3.3.2 Experimental setup . . . 48
3.3.3 Comparison with SOTA and baseline methods . . . 49
3.3.4 Cross-dataset evaluation . . . 53
3.4 Discussion and conclusions . . . 55

4 Temporal Consistency: Long-term 2D and 3D tracking 57
4.1 Problem definition . . . 58
4.2 Methodology . . . 59
4.2.1 Initial processing . . . 61
4.2.2 Model selection . . . 64
4.2.3 Optimisation . . . 65
4.3 Experiments . . . 68
4.3.1 Datasets . . . 68
4.3.2 Evaluation metrics . . . 69
4.3.3 Implementation details . . . 70
4.3.4 Results . . . 72
4.3.5 Ablation . . . 80
4.4 Discussion and conclusion . . . 83

5 Geometric Grounding 85
5.1 Problem definition . . . 86
5.2 Methods . . . 87
5.2.1 Setup . . . 87
5.2.2 Calibration . . . 87
5.2.3 Acquisition protocol . . . 89
5.2.4 Data Processing . . . 90
5.3 Data Records . . . 93
5.4 Technical Validation . . . 95

5.4.1 Calibration 95
5.4.2 Benchmark of state-of-the-art methods 95
5.5 Discussion and Conclusion 109

6 Conclusions **111**
6.1 Contributions 112
6.2 Limitations 114
6.2.1 Dependence on Frame-Level Supervision and Limited Temporal Context 114
6.2.2 Assumptions in Canonical and Implicit Representations 114
6.2.3 Dataset Scope and Generalisation 115
6.2.4 Computational Considerations 115
6.3 Future Research Directions 115
6.3.1 Semantic–Temporal Integration Across Perceptual Representations 116
6.3.2 Adaptive and Topology-Aware Scene Representations 116
6.3.3 Efficient and Deployable Visual Perceptual Systems 116
6.4 Final Remarks 117

A Complementary material Ch. 3 **136**

B Complementary material Ch. 4 **138**
B.1 Experiments on complementary datasets 138
B.1.1 Dataset 138
B.1.2 Implementation remarks 139
B.1.3 Results 139

List of Figures

1.1 Examples of Minimally Invasive Procedures . . . 3
1.2 Robotic Minimally Invasive Surgical systems . . . 6
1.3 da Vinci Research Kit components . . . 7
1.4 Example of Endoscopic video analysis framework . . . 10

2.1 Visual Surgical Perception Framework . . . 19
2.2 Identification of objects in MIP . . . 21
2.3 Tissue segmentation . . . 26
2.4 Representation tracking in MIP . . . 28
2.5 Metric Spatial Representations . . . 34

3.1 DTX network architecture . . . 42
3.2 3D Positional Encoder . . . 43
3.3 Bipartite graph . . . 45
3.4 Contrastive object representations . . . 47
3.5 Qualitative results - Tool detection . . . 51
3.6 Object representation space . . . 52

4.1 InvNeRF framework . . . 60
4.2 Workspace construction . . . 63
4.3 Tool segmentation . . . 63
4.4 Multi-scale HexPlane . . . 64
4.5 Multi-scale error grid sampling . . . 66
4.6 Error map comparison during optimisation . . . 73
4.7 Qualitative depth comparisons of test-time optimisation SOTA . . . 75
4.8 Qualitative results over the STIR and SCARED datasets . . . 78
4.9 Qualitative results over the SurgT dataset . . . 80
4.10 Proposed Invertible Architectures . . . 81

5.1 Data collection setup . . . 88
5.2 3D noise filter . . . 91

5.3 Occlusion masks 92
5.4 Data collection transformation pipeline 92
5.5 Denoising Ex-Vivo image 93
5.6 Data folder structure in DRENDS 93
5.7 Example of the data related to a single time step in a sequence. . . . 94
5.8 Benchmarking: DRENDS-Phantom 102
5.9 Benchmarking: DRENDS-ExVivo 103
5.10 Qualitative depth estimation: DRENDS-Phantom 105
5.11 Qualitative depth estimation: DRENDS-ExVivo 106
5.12 Qualitative depth estimation: models tailored for surgical data 107

A.1 Tools attention maps 137

B.1 Qualitative results over the SimCol dataset 140
B.2 Reconstruction from InvNeRF representation 141

List of Tables

1.1 Levels of autonomy in surgical robotics 8
1.2 Data types in Robotic Minimally Invasive Surgery (RMIS) 9

2.1 Publicly available dataset for depth estimation in surgery 37

3.1 DTX tool detector comparison against SOTA 50
3.2 Comparison of model accuracy broken down by class 51
3.3 DTX ablation paired differences 53
3.4 Cross-dataset evaluation on Cholec80-Boxes 54
3.5 Cross-dataset evaluation per class 54

4.1 InvNeRF tracking comparison against test time optimisation SOTA . 74
4.2 InvNeRF 2D tracking comparison against feedforward SOTA 76
4.3 InvNeRF 3D tracking comparison against feedforward SOTA 77
4.4 InvNeRF comparison on the SurgT validation subset 79
4.5 InvNeRF results per case on the SurgT validation subset 79
4.6 InvNeRF ablation study 80
4.7 Quantitative results for tested architectures 82

5.1 DRENDS camera system calibration results 95
5.2 DRENDS-Phantom benchmark results 102
5.3 DRENDS-ExVivo benchmark results 103
5.4 DRENDS benchmark results: models tailored for surgical data 108

A.1 DTX hyper-parameters search and ablation study 136

B.1 Quantitative results for SimCol dataset 140

Glossary

Artificial Intelligence (AI) A field of computer science concerned with the development of systems capable of performing tasks that typically require human intelligence, such as perception, reasoning, learning, and decision-making.

Computer Vision A field of computer science concerned with enabling machines to acquire, process, and interpret visual data to extract meaningful information about the environment.

Computer-Assisted Interventions (CAI) Clinical interventions supported by computational systems that provide assistance through planning, guidance, visualisation, or decision support during medical procedures.

Computer-Assisted Surgery (CAS) A set of computer-assisted interventions focused on surgical procedures, where computational methods support the surgeon through image guidance, navigation, or intraoperative assistance.

Endoscopy Image modality in which an endoscope is used to visualise internal organs or cavities for diagnostic, screening, or therapeutic purposes.

Hand-Eye Calibration Estimation of the rigid transformation relating the coordinate system of a robot's end effector to that of a camera mounted on it, without which kinematic and image measurements cannot be expressed in a common coordinate system.

Image Understanding The process of extracting, organising, and interpreting semantic information from images by transforming low-level visual data into higher-level representations.

Minimally Invasive Procedure Clinical procedure that is performed through small incisions or natural orifices using specialised instruments and image guidance to reduce patient trauma.

Minimally Invasive Surgery A surgical approach in which procedures are performed through small incisions using specialised instruments and image guidance to reduce patient trauma..

Object Representation A structured encoding of an object's properties, such as appearance, geometry, position, or state, used to support recognition, tracking, or reasoning within a computational system.

Robotic Minimally Invasive Surgery (RMIS) Minimally invasive surgical procedures in which robotic systems mediate the interaction between the surgeon and the patient to enhance dexterity, precision, and control.

Robotic-Assisted Interventions (RAI) or Robotic-Assisted Surgery (RAS) Clinical interventions in which robotic systems assist the execution of a procedure by enhancing precision, dexterity, or control under human supervision.

Scene Understanding The process of constructing a coherent semantic and spatial representation of a scene by integrating information about objects, geometry, relationships, and dynamics.

Surgical Data Science (SDS) An interdisciplinary field that studies how data generated before, during, and after surgery can be acquired, analysed, and modelled to improve surgical understanding, decision-making, and outcomes.

Surgical Perception The computational process through which surgical data are transformed into structured representations of the surgical environment to support interpretation, reasoning, and action.

Abbreviations

2D Bidimensional.

3D Three-dimensional.

3DGS 3D Gaussian Splatting.

aabb Axis-Aligned Bounding Box.

AI Artificial Intelligence.

CAI Computer-Assisted Interventions.

CAS Computer-Assisted Surgery.

CL Contrastive loss.

CT Computed Tomography.

CV Computer Vision.

dVRK da Vinci Research Kit.

FoV Field of View.

FPS Frames per Second.

GIoU Generalised Intersection Over Union.

GT Ground Truth.

HPC High-Performance Computing.

ICG Indocyanine Green.

IoU Intersection Over Union.

IR Infrared.

MIP Minimally Invasive Procedure.

MIS Minimally Invasive Surgery.

NeRF Neural Radiance Fields.

PSNR Peak Signal-to-Noise Ratio.

RAI Robotic-Assisted Interventions.

RAS Robotic-Assisted Surgery.

RGB Red, Green, and Blue.

RMIS Robotic Minimally Invasive Surgery.

ROS Robot Operating System.

SDS Surgical Data Science.

SfM Structure from Motion.

SLAM Simultaneous Localisation and Mapping.

SSIM Structural Similarity Index.

ToF Time-of-Flight.

Tx-decoder Transformer decoder.

Tx-encoder Transformer encoder.

WHO World Health Organization.

Chapter 1

Introduction

Surgical procedures play a crucial role in diagnosing and treating diseases. According to the World Health Organization (WHO), more than one surgical procedure is performed each year for every 25 people worldwide, and the trend also suggests a growing demand as technology evolves and people's longevity increases [1]. In the United Kingdom, surgical procedures are organised by anatomical systems following the national standard Classification of Interventions and Procedures (OPCS-4) [2], supporting an anatomy-driven stratification that reflects clinical differences between procedures. In particular, procedures in the abdominal cavity stand out since a large fraction of the surgical needs worldwide is related to organ diseases in this area [3]. The importance of abdominal surgery, along with the advantages that new technologies such as robotics and imaging through endoscopy provide to the surgeon and patient by reducing invasiveness or physiological burden [4], [5], defines the clinical scope of this work. The incorporation of Artificial Intelligence (AI) into the processing of surgical-related data [6] then determines the methodological approach taken within that clinical scope. Hence, this work focuses on analysing and modelling surgical data from procedures involving the visualisation of organs in the abdominal area through an endoscopic camera using AI techniques. To expand the clinical context, specify the research area, and define the gap addressed in this thesis, a brief background on minimally invasive procedures, Computer-Assisted Surgery (CAS), robotic surgery, surgical data science and surgical scene perception is presented.

1.1 Minimally Invasive Procedures (MIP)

Minimally Invasive Procedure (MIP) is today's gold standard in the surgical room for the diagnosis and treatment of diseases inside the human body [7], and endoscopy

plays an essential role as the main enabling technology [8]. They have brought the best outcomes for patients compared to traditional open surgery by reducing the detrimental side effects of surgeries, reducing pain and recovery times, and increasing survival rates and life expectancy [9], [10]. MIP and endoscopy have a long and intertwined history [11], ranging from antique surgical practices for viewing body cavities [12] to the development of new technologies that enable high-quality visualisation [13], and to the incorporation of robotics and enhanced imaging [14]. In this work, MIP refers to a surgical or diagnostic approach that is performed through small incisions or natural orifices, using specialised instruments and image guidance to reduce patient trauma[15]. Meanwhile, endoscopy refers to the image modality in which an endoscope (laparoscope, when it is used in the abdominal area), i.e., a thin tube-like instrument with a light and a lens for viewing, is used to visualise internal organs or cavities for diagnostic, screening, or therapeutic purposes, independent of whether surgery is performed. Other imaging modalities, such as ultrasound, radiography and magnetic resonance, that can support MIP are acknowledged. However, these are left out of this work scope since this work is concerned with the procedures in which the operative scene is observed through an endoscopic camera. Thus, the procedures studied in this thesis are laparoscopic, performed either manually or with robotic assistance.

Surgical MIP is denominated as Minimally Invasive Surgery (MIS), and it is carried out with therapeutic intent, implying the treatment of tissues through cutting and manipulation. When it is executed over the abdominal area, it is often referred to as a laparoscopic procedure [16]. The general steps for a laparoscopy are as follows [17]: the abdomen is inflated with CO_2 gas to generate a cavity before several small incisions are made through the abdominal wall. Then, the laparoscope (a rigid long endoscope that stays static while surgical tools move) is inserted through one opening to guide the surgery, and other tiny surgical instruments are inserted through other openings, allowing a surgeon to manipulate the patient's organs intracorporeally and perform the surgery. MIS requirements [16] involve that the organs of interest can be visualised through a set of incisions in the body, there is enough room to manipulate the tissue (considering the size of the tools and the movements that will be executed), and the patient presents a homeostatic stable state (i.e. a relatively constant balance of the human body's internal variables).

Diagnostic MIP involves the insertion of the endoscope to screen organs (using a semi-rigid or flexible endoscope to facilitate the displacement of the camera through the body) with minimal or no intervention of other surgical instruments, and is performed to establish or confirm a diagnosis rather than to treat [18]. The clinical distinction between diagnostic and surgical MIP rests on intent rather than on access

route, and the two do not partition cleanly: diagnostic laparoscopy reaches the abdominal cavity through incisions and may proceed to therapeutic intervention within the same session [19]. The property that matters for the analysis of the images in this thesis is therefore not the intent of the procedure but the behaviour of the endoscope, whether it constantly moves or remains mostly static: an endoscope that translates along a lumen, as in colonoscopy or gastroscopy, against a laparoscope that pivots about a fixed port. Examples of diagnostic MIP that involve the abdominal area are colonoscopy, where the endoscope enters through the anus, travels the rectum and colon; and gastroscopy, where the endoscope enters the mouth and travels the oesophagus, stomach, and duodenum.

Figure 1.1 shows a representation of each MIP discussed in this section and highlights the difference in the behaviour of the endoscope between surgical and diagnostic MIP. While most of the work in this thesis focuses on abdominal MIP, the constant camera movement in diagnostic MIP adds an extra degree of complexity to the analysis of the endoscopic images.

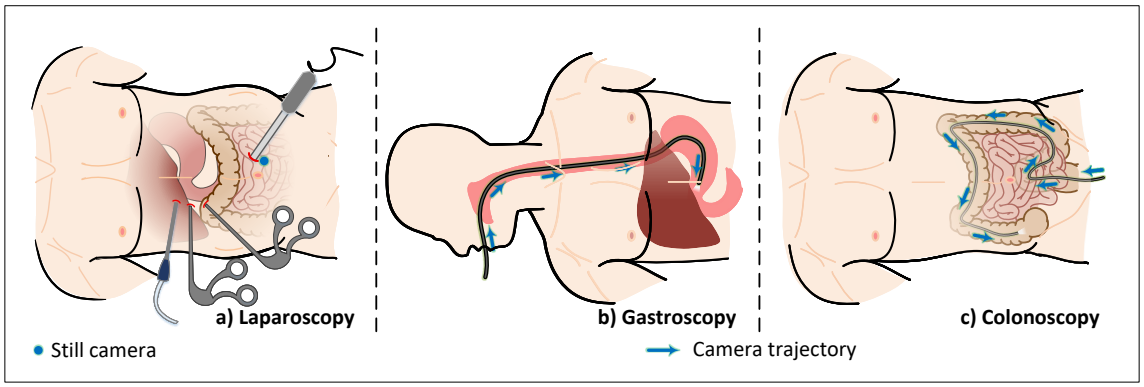


Figure 1.1: **Examples of MIP:** a) Laparoscopy, b) Gastroscopy, c) Colonoscopy.

1.2 Robotic MIP

Significant efforts and emerging trends in the analysis of data from MIP aim to support surgeons in the operating room, ultimately leading to the development of techniques that leverage the capabilities of modern computational systems [6]. Additionally, numerous attempts to increase the complexity of the instruments' movements, precision, and range of motion led to the development of surgical robotic platforms [20]. In this context, endoscopic imaging remains the primary source of information about the operative scene [21], while additional sources such as the robot state and system constraints provide new structured data [20]. As a result, robotic MIP offers a natural framework for studying how visual information can be

interpreted and exploited to enhance diagnostic and surgical performance, safety, and autonomy. This section introduces the motivation of this thesis through the role of computer assistance and robotic systems in MIP, with particular emphasis on how endoscopic image analysis underpins higher-level capabilities.

1.2.1 Computer-Assisted Interventions

Computer-Assisted Interventions (CAI) refers to implementations in which computational systems support or augment interventional procedures, typically by providing intraoperative guidance, visualisation, navigation, or execution support to the surgical team [20], [22]. In the literature, CAI is often referred to as CAS while maintaining in the scope both diagnostic and surgical procedures [6]. Hence, we use both terms interchangeably in this work.

Representative examples of CAI systems are drawn from established surveys in the Handbook of Robotic and Image-Guided Surgery [23]. These span different surgical domains: Intuitive Surgical and Medtronic operate in laparoscopic surgery, whereas Brainlab, Stryker and Zimmer Biomet address cranial and orthopaedic procedures, where the anatomy is rigid and the imaging and guidance requirements differ accordingly. For instance, Intuitive Surgical provides solutions that support planning, intraoperative guidance, and execution by integrating endoscopic imaging, instrument kinematics, and workflow software within a unified robotic surgical platform; Medtronic develops technologies that assist preoperative planning and intraoperative navigation by combining imaging, tracking, and robotic execution within minimally invasive procedures; Brainlab offers systems focused on surgical planning and navigation, leveraging imaging data and intraoperative guidance to support precise spatial decision-making; Stryker provides platforms that assist procedure-specific planning and guidance by integrating preoperative models, intraoperative sensing, and constraint-based execution support; Zimmer Biomet applies techniques to support surgical planning and execution, using data-driven guidance and robotic assistance to improve procedural consistency, particularly in orthopaedic contexts. The degree to which each of these capabilities has reached the operating room varies, and several remain at the stage of development or limited clinical evaluation rather than routine use. Across these CAI systems, a common trend emerges: the progressive integration of perception, planning, and execution within unified computational pipelines, with robotic platforms increasingly serving as the interface between algorithmic decision-making and physical action. CAI is not, however, confined to robotic platforms. Computer-vision assistance is also deployed in manual laparoscopy, as in the Maestro system, whose ScoPilot capability

automates laparoscope control on conventional laparoscopic instrumentation [24], [25].

Regarding MIP, it represents the primary clinical domain in which CAI technologies deliver tangible value, as they fundamentally alter how surgeons interact with the surgical environment [23], [26]. The reduced access, indirect visualisation, and constrained instrument motion characteristic of MIP create intrinsic challenges (steep learning curve, camera limitations and scene dynamics) that cannot be easily addressed by human skill alone [16], [27]. CAI systems, such as navigation platforms, robotic assistants, and intraoperative guidance tools, are therefore developed to compensate for these limitations by enhancing spatial awareness, precision, and reproducibility.

1.2.2 Robotic platforms and levels of autonomy

Surgical robotics, often referred to as Robotic-Assisted Interventions (RAI) or Robotic-Assisted Surgery (RAS), represents a further step in the integration of CAI in the operating room by introducing robotic manipulators that physically interact with the patient [28]. The use of Robotic Minimally Invasive Surgery (RMIS) is increasingly adopted to improve dexterity, precision, and ergonomics in MIP, while preserving the reliance on endoscopic visual feedback [20]. The da Vinci Surgical System, Hugo RAS, Versius, and Senhance (Figure 1.2) are examples of prominent robotic platforms designed to support RMIS through teleoperated or computer-assisted manipulation [23]. The da Vinci system combines articulated instruments with stereoscopic endoscopic vision to enable precise manipulation in confined operative spaces; Hugo RAS adopts a modular architecture with independently positioned robotic arms and a digital workflow to integrate with standard operating rooms; Versius employs multiple bedside robotic arms and an open surgeon console to support flexible port placement and ergonomic control in laparoscopic procedures; and Senhance augments conventional laparoscopy with robotic assistance by providing instrument articulation, camera control, and haptic feedback while preserving a laparoscopic-style operating environment.

Among all systems, the da Vinci Surgical System has stood out as the most widely used platform for research due to its openness and reproducibility [5]. Johns Hopkins University, Worcester Polytechnic Institute, and Intuitive Surgical, Inc. developed the da Vinci Research Kit (dVRK) [33] (Figure 1.3 shows the main components of the dVRK platform). This kit provides the system's core hardware and dedicated electronics, enabling control at any level, thus driving research trends in RMIS, both for industry and academia.

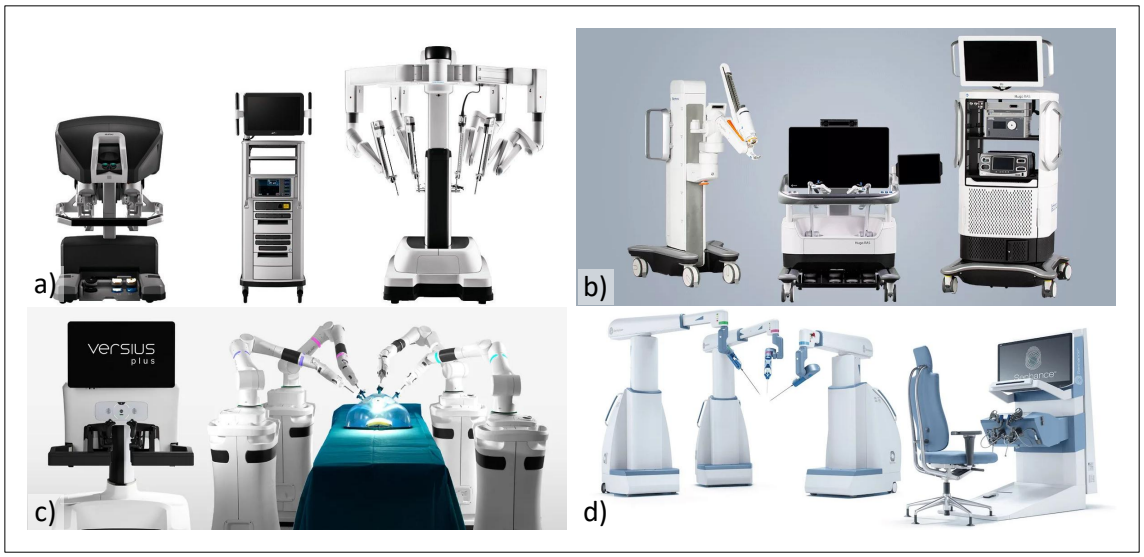


Figure 1.2: **Examples of RMIS:** a) da Vinci Surgical System - Intuitive Surgical [29], b) Hugo RAS System - Medtronic [30], c) Versius Surgical System - CMR Surgical [31], d) Senhance Surgical System - Asensus Surgical [32].

Despite the technological sophistication of current RMIS platforms, randomised controlled trials comparing robotic-assisted and laparoscopic surgery in rectal cancer resection have shown no significant reduction in conversion to open laparotomy when using robotic systems, even across surgeons with varying levels of robotic experience [34]. Similarly, the ROC'N'ROLL trial comparing robotic and laparoscopic hepatectomy for liver malignancies reported comparable short-term outcomes, including complication rates and operative metrics [35]. These findings suggest that enhanced instrument articulation and improved ergonomics alone are insufficient to guarantee superior clinical performance. As highlighted in recent analyses of medical robotics translation, robotic systems are insufficiently integrated with real-time intraoperative imaging and clinical workflow [36]. This limitation may be attributed not to mechanical capability but to perceptual capability: integrating imaging into the intraoperative workflow requires a system that can interpret what that imaging shows, and current platforms do not. Robotic manipulators enhance dexterity and precision, and open the opportunity for new forms of computer-assisted intervention, but they do not fundamentally augment the system's understanding of the operative scene, which is the capability this thesis sets out to build.

Beyond supporting human decision-making and teleoperation, recent trends in surgical robotics explore varying degrees of automation [37]. An autonomous robotic system is a sophisticated and advanced technology that can navigate its environment, perceive and interpret sensory data, plan and execute actions, and adapt to

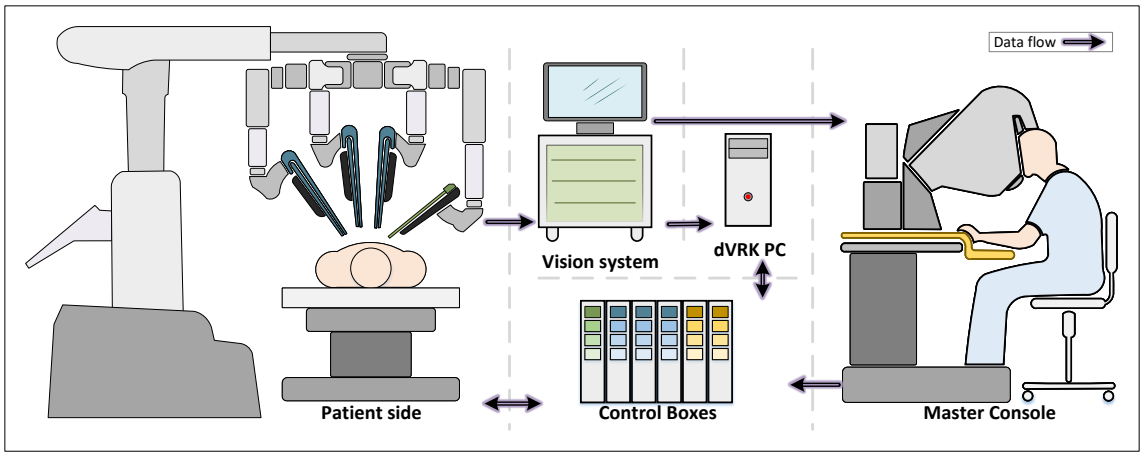


Figure 1.3: **dVRK platform components.** Patient side is near the operating table and holds robotic arms with surgical instruments. Robotic arms are controlled through commands from the master console or dVRK PC. The Control Boxes house the electronics that control the robot and translate commands into robot movements. The master console consists of hand controls, foot pedals, and a viewing screen used by the operator. Vision systems hold the electronics to receive and transmit data from the endoscopic camera. dVRK PC hosts the research software to communicate with the system.

changing conditions, all without direct human control [20], [37]. In a surgical context, tasks such as suturing and knot tying [38], tissue dissection [39], instrument reconfiguration [40], or organ manipulation [41] have directed the research field. Table 1.1 shows a classification for the levels of autonomy in RMIS and their characteristics. The automation of surgical tasks has been mainly conducted under two paradigms [42]: automation informed by machine learning and automation by reinforcement learning. The automation informed by machine learning leverages data to identify patterns and apply them in decision-making [43]. In contrast, automation by reinforcement allows the system to learn from interaction with the environment (which can be virtual) by providing feedback in the form of rewards or penalties [44]. The mechanisms and data types in each paradigm differ, but they are not necessarily mutually exclusive, and both optimise their decision-making strategies based on interpretations of the surgical environment. However, level 0 remains the only form of automation used in clinical practice, as the work in this field has yet to provide a complete solution that addresses the requirements of a real surgical setup [37]. The development of such capabilities is fundamentally constrained by the robot's ability to perceive and interpret its environment [4], [6], [37], [43], including anatomical structures, surgical instruments, and tissue behaviour. In this context, endoscopic image analysis plays a critical role, providing the visual perceptual foundation upon which more advanced robotic assistance and autonomy can be built [21].

In summary, robotic MIP have substantially advanced surgical practice by enhancing dexterity, precision, and ergonomics. However, current clinical evidence indicates that these mechanical advantages do not consistently translate into superior patient outcomes compared to conventional laparoscopy. This gap suggests that the next transformative step in robotic surgery will not be driven solely by hardware innovation, but by the development of robust perceptual capabilities that enable structured understanding of the surgical scene. The integration of endoscopic imaging, instrument kinematics, and new surgical workflows in these platforms provides the foundations for enhanced guidance, reproducibility, and emerging forms of automation. Hence, the analysis of surgical data emerges as a central enabling component for overcoming current limitations in robotic MIP, thereby motivating this thesis's investigation into surgical perception as the pathway toward more informed, robust, and adaptive MIP systems.

Table 1.1: General description of the levels of automation in robotic minimally invasive surgical systems [37]

Autonomy level	**Description**
Level 0	No autonomy, all the movements of the robot are controlled by the surgeon and no support or constraints are provided.
Level 1	Autonomy for assistance. The robot autonomously provides enhanced feedback to the surgeon.
Level 2	Task autonomy, the execution of certain tasks is autonomously performed by the robot once the surgeon has provided all the needed information to accomplish the task.
Level 3	Conditional Autonomy, the robot has a perceptual capability to understand the surgical scene and autonomously plan and execute specific tasks when the surgeon requires assistance.
Level 4	High autonomy, the robot uses pre- and intra-operative information to plan and execute a sequence of tasks (plus replanning if needed). The surgeon can re-take control at any point.
Level 5	Full automation, the robot has complete control to analyse the pre- and intra-operative information, devise a plan and perform the surgery without human intervention. A platform with this characteristic is far from the current state of the art, and proper ethical discussion needs to be addressed to define their constraints

Table 1.2: Available data types during a Robotic Minimally Invasive Procedure [20].

Data type	Description
Preoperative data	All available patient data acquired before the operation, such as blood tests, radiography or any other piece of information that could be in an electronic health record.
System data	Teleoperation states: pedal signals, head sensor, presence sensor, etc.
Kinematic and Dynamic data	All the information related to the manipulators and actuators of the system on the console and patient sides.
External data	Any additional sensors such as ultrasound probe or other non-invasive imaging devices.
Endoscopic images	Frames from the stereo endoscope or any other camera.

1.3 Surgical Data Science

Surgical Data Science (SDS) has emerged in the last decade as a discipline that studies and exploits data generated throughout the surgical workflow (pre-, intra-, and post-operative) [45], with the aim of extracting actionable knowledge by organising, analysing, and modelling such data to improve surgical understanding, decision-making, and outcomes [6].

Table 1.2 displays the data types available during a robotic MIP [20]. Among these, endoscopic images represent the dominant source of intraoperative information [21] and are often the only modality providing direct and continuous visualisation of the surgical field [45]. In contrast, robot kinematics, dynamics, and system data are explicitly designed to represent the state of the two agents involved in the procedure, the robot and the surgeon, in a structured and measurable manner, with the console-side and patient-side manipulators being the components through which their state is measured [20]. Patient-related data during surgery are largely limited to vital signs [16], which provide global physiological status but little information about the local surgical scene, while pre-operative data primarily support planning and post-operative assessment rather than real-time intervention [22]. Finally, other imaging modalities are used to support the interpretation of endoscopic images or, when serving as the primary visual sensor, are significantly less widely adopted across procedures than endoscopic vision [21]. Given the relevance of endoscopic images intraoperatively and their lower abstraction level compared to other data types, a principled notion of vision-centred perception is required to transform rich, unstructured visual sensory data into representations suitable for analysis, reasoning and decision-making.

1.3.1 Visual Surgical Perception

In this thesis, perception is grounded in the Computer Vision (CV) ideas proposed by David Marr [46], in which visual perception is understood as a set of processes that construct representations from visual data to support increasingly complex forms of reasoning. Hence, surgical perception is defined as the computational process through which surgical data, mainly visual, are transformed into structured representations of the surgical environment that support understanding, reasoning, and action. Although this definition admits every form of surgical data, this thesis takes endoscopic video as the input to perception and treats the remaining sources, such as robot kinematics and system state, as context that constrains or supports its interpretation. Within this framework, visual surgical perception refers not merely to recognising visual patterns in endoscopic images, but to constructing spatially and temporally consistent representations of the operative environment that capture what is present, where it is located, how it moves, and how different elements interact. Such understanding is essential to enable safer decision-making, effective human–robot interaction, and progress toward higher levels of automation [47]. Consequently, visual surgical perception defines the central study object of this thesis, motivated by advancing robotic MIP toward more informed, adaptive, and robust interventions.

The concept of visual surgical perception is complemented by the definition of image understanding in computer vision, which associates vision tasks with the ex-

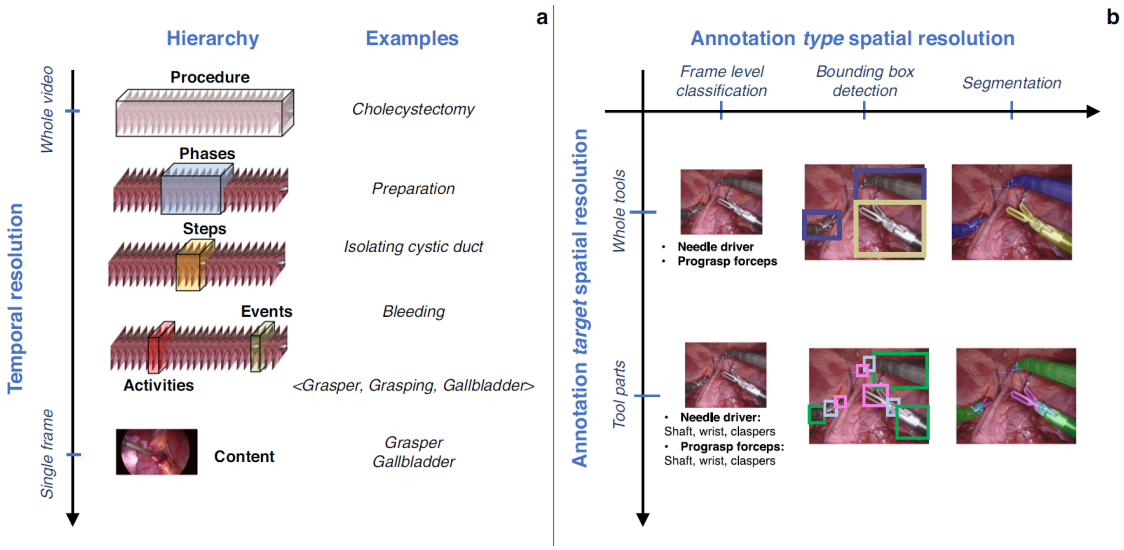


Figure 1.4: **Endoscopic video analysis framework** [48].a) Different temporal resolutions of the data in a surgical procedure. b) Different label resolutions associated with different computer vision tasks.

traction and organisation of semantic information from images, depicting different levels of abstraction by transforming low-level visual data into progressively higher-level representations [49]. In a surgical context, efforts have been made to classify visual tasks on surgical video analysis that support CAI systems, and to depict the dependency that exists among them [45], [50]. For example, Mascagni [48] proposed a framework for the characterisation of surgical tools in MIS where temporal video resolutions and label granularity are linked to different levels of surgical understanding (Figure 1.4). In this example, the approaches for tool identification within the image frame extract representation of the objects that can increase specificity, but still answer the same questions: what and where an object is on the image plane. Only the analysis over temporal windows allows the extraction of new information to answer more complex questions. In traditional computer vision, a hierarchy is underpinned by the dependencies that exist between tasks when aiming to extract a high-level abstraction from visual data, such as describing the interaction between objects, which cannot be achieved without identifying the objects involved in the interaction [49]. Despite notorious advances in the SDS, to the best of my knowledge, there is currently no unified framework that explicitly relates this task hierarchy, levels of assistance for robotic MIP, and the integration of heterogeneous surgical data sources within a single conceptual model. Such a synthesis is particularly relevant for robotic surgery, where different levels of autonomy depends on specific levels of understanding to support spatial reasoning, temporal consistency, and interaction with a dynamic environment [37], [47]. Nevertheless, some works have hinted at this relationship by defining key technologies and visual tasks as enablers of complex CAI methods [45], [47], [50]. For instance, tool detection and tracking have been widely adopted for context-aware assistance [48], depth estimation and camera trajectory estimation are preliminary steps for 3D reconstruction [49], while 3D reconstruction and tissue tracking are increasingly recognised as prerequisites for spatial reasoning and autonomy [20].

Ultimately, the common objective underlying visual perceptual capabilities is not merely the extraction of isolated semantic labels or geometric quantities, but the construction of a unified spatial and temporal reference in which all relevant elements of the surgical scene can be coherently represented and related. In modern surgery, effective decision-making and guidance rely on the ability to express instruments, anatomy, tissue deformation, camera motion, and pre-operative information within a common coordinate framework that enables consistent spatial reasoning and interaction [21], [37]. From this perspective, surgical perception is not an end in itself, but a necessary mechanism for aligning disparate sources of surgical information into a common representational space. The absence of such a coordinated

perceptual framework remains a fundamental limitation of current robotic MIP systems, motivating the problem addressed in this thesis.

1.3.2 Learning-Based Approaches for Visual Perception

The complexity and variability of surgical scenes pose fundamental challenges for visual perception based solely on rule-based or hand-crafted approaches [6]. Endoscopic images acquired during MIP exhibit significant variability in illumination, viewpoint, anatomy, tissue appearance, deformation, and the presence of fluids or occlusions [27]. These factors make it difficult to design deterministic models that generalise reliably across procedures, patients, and surgical contexts [50]. As a result, learning-based methods have emerged as a natural and effective paradigm for addressing visual perception tasks in surgical environments, enabling models to adapt to variability by learning representations directly from data [43]. Within the broader scope of SDS, AI refers to computational methods that enable systems to learn patterns, relationships, and representations from data in order to support interpretation, prediction, and decision-making [47].

The learning-based solutions for computer vision tasks investigated in this thesis are organised according to the scene representations they provide and the reasoning they enable, rather than by specific algorithmic families. Methods for object identification within the image space could enable context awareness and support downstream analysis, such as workflow understanding and human–robot interaction [48]. Beyond static interpretation, methods that integrate temporal information capture motion and continuity over time, which are essential for tracking instruments, tissue deformation, and camera motion in dynamic surgical environments [51]. Finally, geometry-aware approaches are explored to recover spatial structure and depth, enabling Three-dimensional (3D) scene reconstruction and, potentially, geometric reasoning, which are critical for navigation, spatial planning, and autonomous decision-making [52]. Together, these complementary classes of learning-based methods support the construction of temporally consistent, spatially grounded representations suitable for integration into robotic control and higher-level surgical reasoning.

While learning-based approaches currently define the state-of-the-art, they also introduce challenges related to data availability, generalisation, interpretability, and robustness in safety-critical settings [47]. Addressing these challenges requires careful problem formulation, dataset design, and evaluation strategies tailored to the surgical domain [6]. For this reason, this thesis focuses on the development and analysis of learning-based perception methods grounded in endoscopic imagery, intending to advance scene representations as a foundation for more informed and

capable computer-assisted and robotic surgical systems. Detailed descriptions of the specific models, training strategies, and experimental settings employed in this work are provided in the subsequent chapters.

1.4 Visual Surgical Perception Limitations

Despite significant advances in robotic MIP systems, current surgical technologies remain limited in their ability to construct a coherent and comprehensive understanding of the operative scene from the surgical data. Surgical procedures generate heterogeneous data streams, ranging from endoscopic images and robotic kinematics to system states and pre-operative information, that exist at different levels of abstraction and are not inherently aligned in space or time. While each data source provides valuable partial observations of the scene, endoscopic images remain the primary intraoperative source of information, providing an unstructured representation of the scene that cannot be easily integrated across time or with other data types. Without a framework that clearly maps modern surgery requirements and visual perception capabilities, key elements of the surgical scene, such as instruments, anatomy, tissue motion, camera motion, and spatial structure, are often analysed in isolation, limiting the scalability, robustness, and autonomy of robotic MIP systems. Addressing this fragmentation requires visual perception methods capable of transforming raw intraoperative endoscopic images into structured, spatially and temporally consistent representations that can serve as a common reference for scene understanding, decision-making, data integration and reasoning.

1.5 Aim of This Thesis

The aim of this thesis is to advance the understanding of dynamic surgical scenes by constructing spatially and temporally coherent representations from endoscopic imagery. Rather than addressing surgical perception in its broadest sense, this work focuses on visual perception to know how information can be progressively structured and integrated to provide a unified representation of the operative scene that supports reasoning about what is present, how it evolves over time, and where it is located in the physical world. Research questions (RQ) that guided this thesis:

RQa. - *How do computer vision tasks, through the representations they produce, define the limitations of visual surgical perception?*

RQb. - *How can instruments and tissues be represented so that their identity is recov-*

ered reliably under scale variation and visual similarity, at the rate an intra-operative system requires?

RQc. - *How can correspondences on deforming tissue be maintained across a whole sequence rather than between neighbouring frames, and how can camera motion be separated from tissue motion?*

RQd. - *What is required to measure metric depth accuracy and temporal consistency in deformable surgical scenes, and how do current models perform when measured that way?*

1.5.1 Objectives

This thesis investigates three complementary levels of understanding. At the semantic level, surgical elements are identified and localised within the image to define the instruments and tissues present in the scene. At the temporal level, long-term correspondences are established to model motion and continuity in dynamic and deformable environments. At the geometric level, spatial structure is recovered to ground these representations in three-dimensional space. Together, these levels enable the construction of a coherent representation of the surgical scenario, forming the basis for scene understanding in robotic and computer-assisted interventions. The relationship between these capabilities, scene understanding, and reasoning-enabling is described further in Chapter 2, together with the families of computer vision methods associated with each capability.

Identification of surgical instruments and tissues. To develop methods for the robust identification and localisation of surgical instruments and tissues in endoscopic imagery, enabling the construction of meaningful representations of the elements in the scene.

Temporal consistency tracking. To model the temporal evolution of the surgical scene through dense and temporally consistent two- and three-dimensional tracking of instruments and tissue over extended sequences.

Metric spatial grounding. To address key limitations in the development and evaluation of depth estimation and three-dimensional reconstruction methods for deformable surgical scenarios, enabling the recovery of metric spatial information from endoscopic data.

1.6 Contributions

This thesis contributes a coherent, perception-driven framework for advancing visual scene understanding in robotic MIP, grounded in the progressive development of perceptual capabilities from endoscopic imagery, presented in Chapter 2, which orders the computer vision tasks of surgical perception by the representations they produce and locates the limitations of the field at the couplings between them rather than within the individual tasks. It introduces methods for real-time detection of surgical tools, temporally consistent two- and three-dimensional tracking of instruments and deforming tissue, and metric depth estimation in visually and geometrically challenging surgical environments. In addition, the thesis presents a novel dataset for depth estimation in robotic endoscopy under tissue deformation, together with open-source processing pipelines and benchmark evaluations. Collectively, these contributions support the construction of unified spatial–temporal representations of the surgical scene.

1.6.1 Publications Arising from This Thesis

- **Loza, G.**, Valdastri, P., Ali, S. (2024). Real-time surgical tool detection with multi-scale positional encoding and contrastive learning. Healthc. Technol. Lett. 11, 48–58. **(published)**
- **Loza, G.**, Hu, J., Jones, D., Ali, S., Valdastri, P. (2025). Surg-InvNeRF: Invertible NeRF for 3D tracking and reconstruction in surgical vision.

  https://www.arxiv.org/pdf/2508.09681. **(under review)**
- **Loza, G.**, Magro, M., Calme, B., Hu, J., Jones, D., Ruffaldi, E., De Momi, E., Ali, S., Valdastri, P. (2025). DRENDS: A Dataset for Depth in Robotic Endoscopy with Dynamic Scenarios. **(under review)**

1.6.2 Relevant Collaborations

- Zeng, T., **Loza, G.**, Hu, J., Valdastri, P., Jones, D. (2024). Realistic Surgical Image Dataset Generation Based on 3D Gaussian Splatting. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol 15006. Springer, Cham. **(published)**
- Zeng, T., Hu, J., **Loza, G.**, Ali, S., Sarikaya, D., Valdastri, P., Jones, D., (2025). NeeCo: Image Synthesis of Novel Instrument States Based on Dynamic

and Deformable 3D Gaussian Reconstruction. IEEE Transactions on Medical Imaging **(published)**

- Magro, M., Covallero, N., Gambaro, E., **Loza, G.**, Jones, D., Ruffaldi, E., Valdastri, P., De Momi, E. (2026). Self-Supervised Metric Depth Estimation for Robotic Surgery Using LoRA-Adapted Vision Transformers. **(under review)**

- Hu, J., Jones, D., **Loza, G.**, Huang, S., Valdastri, P., (2025) Autonomous Robotic Exploration of Unknown Soft Objects. The International Journal of Robotics Research. **(published)**

- Calmé, B., Greenidge, N. J., Metcalf, A., Bacchetti, A., **Loza, G.**, Kpeglo, D., Lloyd, P., Chandler, J. H., Pensabene, V., Valdastri, P. Moving Beyond Compliance in Soft-Robotic Catheters Through Modularity for Precision Therapies. **(under review)**

1.7 Thesis Structure

This thesis is organised around the specific objectives.

Chapter 2 provides the technical background required for the remainder of the thesis. It introduces the foundational concepts in computer vision, machine learning, and surgical data science that underpin the visual and surgical perception framework in this thesis, followed by key related computer vision tasks and a deep discussion of state-of-the-art techniques to highlight specific gaps in the field. Together, this chapter establishes the technical vocabulary and context needed to interpret and situate the subsequent chapters. Beyond this background, it contributes the framework that orders those tasks by the representations they produce, on which the organisation of the remaining chapters rests.

Chapter 3 focuses on the identification of surgical instruments in endoscopic images. It addresses surgical tool detection using a transformer-based approach that learns consistent, discriminative embedded representations, enabling real-time localisation and classification of instruments.

Chapter 4 addresses temporal consistency tracking, with an emphasis on long-term tracking. This chapter introduces a new method for tracking based on invertible neural networks embedded within a test-time optimisation framework using Neural Radiance Fields (NeRF). Results and discussion focus on how temporal coherence supports the interpretation of dynamic surgical scenes.

Chapter 5 focuses on metric spatial grounding, addressing the lack of depth ground truth in dynamic surgical scenes. This chapter introduces a novel dataset capturing depth information in robotic endoscopy under tissue deformation and presents a baseline evaluation.

Chapter 6 concludes the thesis by summarising the contributions, discussing limitations, and outlining directions for future research toward visual perception-driven autonomy in robotic minimally invasive surgery.

Chapter 2

Foundations and Related Work

Visual Surgical perception, as defined in Chapter 1, is a progressive process of representation building, in which increasingly structured and informative descriptions of the operative scene are derived from raw intraoperative endoscopic images. Building on this foundation and using endoscopic imagery as the primary sensory input, representation levels are supported by groups of well-defined CV operations that address complementary questions about the scene: what elements are present, how they evolve over time, where they are located in space, and ultimately how observations can be integrated across viewpoints and temporal scales. This chapter introduces the representation levels adopted in this thesis and reviews the computer vision operations commonly associated with each level, providing the technical background required to contextualise the methodological contributions presented in subsequent chapters.

Figure 2.1 defines the conceptual framework that guides the research presented in this thesis. Each representation level is a step toward a more comprehensive and structured description of the operative scene, while the blocks within each capability denote data or operations that generate, transform, or augment representations with comparable structural properties. The boundaries between levels do not correspond to specific algorithms, but rather to the level of organisation and consistency introduced into the scene. Multiple computational approaches may instantiate the same abstraction level, and a single comprehensive method may span multiple levels.

Within this framework, local primitives refer to simple pixel- or patch-level representations that, in traditional approaches, are often associated with low-level visual features. Identification of instruments and tissues encompasses methods that create object-level representations, providing a semantic anchor for scene understanding. Temporal consistency introduces a temporal anchor, capturing the ability to maintain reliable representations across time steps and scene changes; in minimally

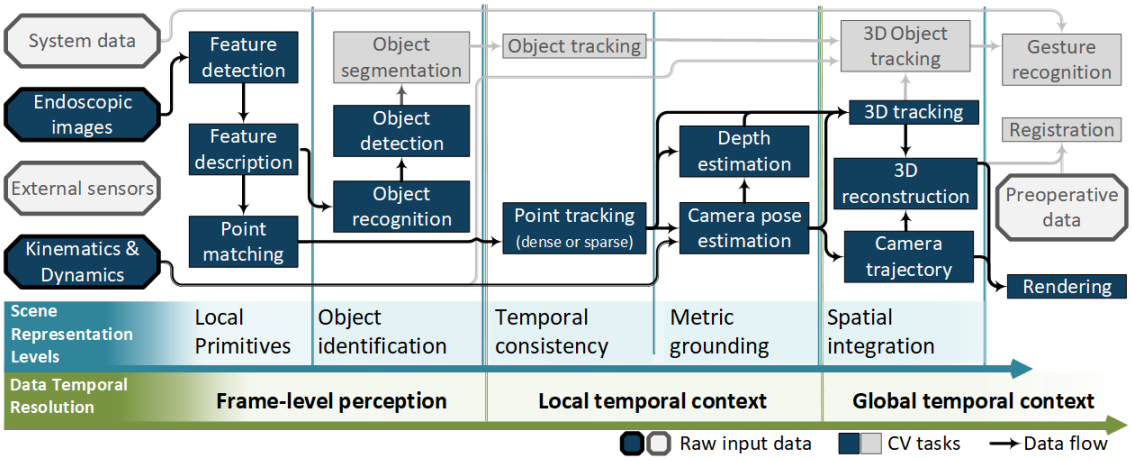


Figure 2.1: **Task-oriented framework for Visual Surgical Perception**. Levels of scene representation are characterised by the types of spatial–temporal reasoning they enable. Within each level, computer vision operations or data (blue and grey blocks) that contribute comparable structural information are grouped, while arrows indicate conceptual dependencies between representations. From left to right, representations progressively increase the semantic, temporal, and spatial organisation of the surgical scene. The band beneath the levels records the temporal extent of the evidence each task requires, which is what places it in one context rather than another: frame-level perception covers tasks decidable from a single image; local temporal context covers those requiring a short window of consecutive frames, such as point tracking, depth from parallax, and camera pose estimation; and global temporal context covers those requiring the whole sequence and several viewpoints, such as camera trajectory estimation and reconstruction.

invasive procedures, this local temporal context can be interpreted as the transition from individual frames to a set of frames capturing surgical actions or events. The introduction of a temporal window further enables metric grounding by allowing tracked representations to be localised in three-dimensional space. Finally, at broader temporal and spatial scales, spatial integration brings representations from multiple time steps and viewpoints into a unified coordinate framework, enabling scene reconstruction.

As discussed in Chapter 1, surgical perception is not solely a vision problem but a multi-data integration challenge. Accordingly, Figure 2.1 also illustrates the relationship of visual perception with additional surgical data modalities beyond endoscopic imagery. Light grey blocks and connections depict operations and relationships that lie outside the primary scope of this thesis but contextualise its contributions within the broader surgical ecosystem. The remainder of this chapter reviews each representation level and discusses the state of the art (SOTA) of the associated CV tasks.

2.1 Local Primitives

Local primitives provide appearance-level representations at the pixel or small-patch scale, serving as the foundation on which higher-level perceptual capabilities operate. While such representations do not, by themselves, yield a meaningful understanding of the surgical scene, they enable subsequent analysis by providing a quantitative and compact description of image content, effectively reducing the dimensionality of the visual data. For this reason, local primitives are treated in this chapter as foundational representations rather than as a stand-alone perceptual capability.

Traditional computer vision approaches construct local primitives by first detecting points of interest, defined as salient regions around specific image locations, and subsequently describing these points using information extracted from a local pixel neighbourhood [49]. A matching stage then establishes correspondences between descriptors, typically comprising both a selection mechanism to determine candidate comparisons and a similarity measure to perform the matching. Among classical methods, gradient-based descriptors such as SIFT [53], SURF [54], and GLOH [55] have been widely adopted. However, models based on these handcrafted representations have been shown to struggle in the challenging conditions of MIP [56]. Learning-based approaches address some of these limitations by optimising feature detection and description jointly from data. Methods such as DELF [57], D2-Net [58], and R2D2 [59] retain the classical detect–describe–match pipeline while replacing handcrafted components with learned representations. More recent approaches further relax this structure by learning correspondences without explicitly supervising detection or description stages; such methods are discussed in Section 2.3, where they are shown to define the current SOTA for feature matching in dynamic surgical scenes.

2.2 Identification of Surgical Objects

Identification of instruments and tissues provides semantic representations that associate visual evidence with discrete surgical scene components, answering the question of what elements are present in the Field of View (FoV). This capability is fundamental in surgery, as reliable localisation and classification of instruments and anatomical structures establish the semantic anchors required for interpretation and interaction. Throughout this chapter the word *object* is used in its computer vision sense, denoting any region of the image that a detector is trained to delineate and label, which in this setting means a surgical instrument or an anatomical structure. Figure 2.1 shows three operations that synthesise information into three representations: object recognition, detection, and segmentation, which all respond to the same question but at different levels of granularity: image, patch, and pixel levels, respectively.

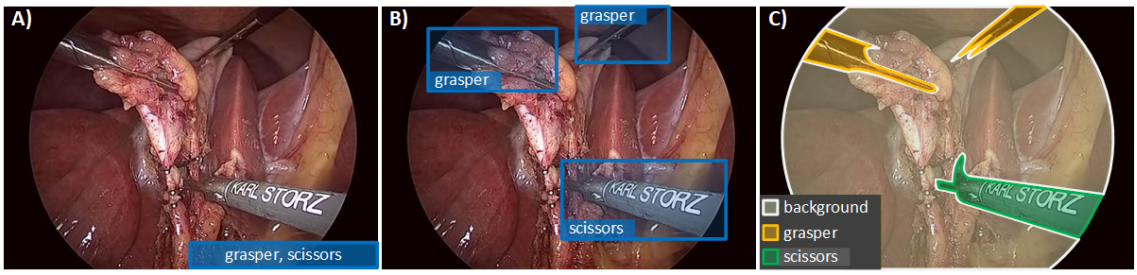


Figure 2.2: **Identification of objects in MIP**. A) Tool recognition, B) tool detection and C) segmentation on an image of the Cholect 80 dataset [60].

Figure 2.2 shows an example of each representation used to identify objects in the FoV; all of them correspond to two-dimensional representations defined on the image plane. Object recognition determines the presence of one or more object categories in an image without explicit spatial localisation. Object detection extends this representation by associating recognised objects with localised regions of the image, typically in the form of bounding boxes. Object segmentation further refines localisation by assigning an object or background label to each pixel in the image.

These representations play a central role in surgical perception, as they enable the identification of instruments and anatomical structures that underpin downstream reasoning, interaction modelling, and skill analysis. Nevertheless, the identification of instruments and tissues alone remains inherently limited: the resulting representations are frame-specific, view-dependent, and lack temporal continuity and metric grounding. As a result, additional visual perceptual capabilities, such as tracking and three-dimensional localisation, are required to integrate semantic information

into coherent, spatially consistent scene representations.

2.2.1 Challenges

In surgical scenarios, challenges associated with the identification of instruments and tissues are exacerbated by the characteristics of intraoperative data and the constraints imposed by MIPs [61]. Significant scale variation, strong class imbalance, and high visual similarity among surgical instruments, combined with low-texture tissue surfaces and continuous forward and backward motion of the endoscopic camera, complicate reliable recognition and localisation [62]. These factors make both multi-class discrimination and scale-aware detection particularly challenging in practice. In addition to visual complexity, the identification of instruments and tissues in surgery is subject to strict operational constraints. Real-time processing is essential, as the interpretation of system and surgeon actions must occur with minimal latency [48], since guidance that arrives after the moment it was needed cannot inform the action it was meant to support. As a result, recognition systems must balance accuracy with computational efficiency, further constraining model design.

At the model level, the transition from handcrafted feature-based approaches to deep learning-based detectors has significantly improved performance [63], but has also introduced new challenges [48], [64]. In particular, receptive field constraints impose a trade-off between capturing fine-grained local features and modelling broader contextual information [62]. Small receptive fields are better suited for distinguishing subtle visual differences between similar tools or tissues, while larger receptive fields facilitate the integration of global context but may obscure local details. In surgical scenes, effective identification of instruments and tissues often requires both levels of representation simultaneously. Another prominent limitation of modern detectors is their reliance on predefined anchor boxes [65]–[67]. Anchors are a fixed set of reference boxes of predetermined scale and aspect ratio, tiled over the feature map, against which the network regresses offsets rather than predicting box coordinates directly; they therefore encode prior assumptions about object size, aspect ratio, and location, which can be detrimental in surgical environments characterised by large variability in object scale, orientation, and appearance [61], [68]. This dependency can reduce robustness and limit generalisation across procedures and imaging conditions.

Finally, data availability remains a critical constraint in surgical perception. The development of solutions for different forms of instrument and tissue representation requires distinct data types and annotation strategies, each associated with specific acquisition challenges [62]. To date, tool-related tasks have benefited from

comparatively more accessible data collection pipelines, enabling earlier validation of complex models [69]. In contrast, the scarcity and cost of annotated data for surgical tissues continue to limit progress and generalisation [70].

2.2.2 Related Work

Tool Identification in MIP

While object recognition, detection, and segmentation all address tool identification, this thesis focuses on object detection as the most appropriate representation for supporting downstream perceptual capabilities in robotic MIP. Detection provides a minimal yet sufficient semantic–spatial description by jointly encoding tool identity and approximate image-plane localisation, enabling temporal association, geometric reasoning, and integration with additional data sources (the camera pose, tool kinematics and geometric priors available in robotic procedures). In contrast, image-level recognition lacks spatial information, while pixel-level segmentation introduces greater representational and computational complexity at early perceptual stages. Tasks whose quantity of interest is a boundary or an area, such as the delineation of anatomical structures and the assessment of the critical view of safety, require full segmentation [71] and are out of the scope of this thesis. In robotic settings, where camera pose, tool kinematics, and geometric priors are often available, detection-based representations can be effectively refined and integrated in three-dimensional space. For these reasons, object detection is adopted as a pragmatic and principled semantic anchor for tools in surgical scene understanding.

The review that follows characterises the design choices adopted across this literature and the limitations they impose. The performance metrics reported in this section comes from the evaluation protocol in each work. However, a controlled comparison under a single protocol is presented in Chapter 3. Early datasets such as ATLAS [72] provided a valuable starting point for proof-of-concept development in robotic MIS under controlled conditions. However, their reliance on mock environments limits their ability to capture the visual variability, motion dynamics, and imaging artefacts encountered in real clinical scenarios, thereby restricting the generalisability of trained models. Real-world data is needed since it is more likely to capture the diversity and complexity of medical cases encountered in clinical practice and produce more robust models whose performances can generalise across patients and hardware [27]. The introduction of spatial annotations in real surgical data marked a turning point for the field. Jin et al. [73] extended the m2cai16-tool dataset by annotating tool locations in 2532 frames, resulting in the m2cai16-tool-

location dataset. Although early detection performance and inference speed were limited (5 Frames per Second (FPS) and 0.6 mAP_{50}), this dataset enabled the first systematic exploration of deep learning–based instrument detection in real surgical scenes and laid the foundation for subsequent SOTA methods.

Research efforts explored both two-stage and single-stage detection paradigms. Zhang *et al.* [74] proposed a Fast RCNN-based model and addressed the problem of anchor dependency with a modulated feature block to incorporate the anchor shape information in the generated feature maps from the backbone. A YOLO-based model was presented by Choi *et al.* [75]. His work reported the fastest inference time of 48 FPS in the m2cai16-tool-location dataset, with a localisation accuracy of 0.722 mAP_{50} over preselected validation videos (Table 3.1). Sai and Sinha [76] presented a multitask model for tool presence, detection (multi-class), and phase classification based on a DSSD architecture (deconvolutional single shot detector). They explained how features from different parts of the architecture can be taken to solve different tasks and achieve improved performance regarding location inference. They did not report inference time, but based on the original Deconvolutional single-shot detector (DSSD) paper [77], the inference time is 15 FPS. Zhao *et al.* [78] proposed a lightweight cascaded CNN architecture from coarse to fine. The first stage in a two-stage detector was similar to a region proposal stage, but with fixed-sized regions. The second was a regression network of the surgical instrument tip region. They reported an inference time of nearly 24 FPS; however, they detected and tracked tip instruments without classification. Similarly, Liu *et al.* [68] proposed a method for tool location without classification over the ATLAS dataset and a relabeled version of the Endovis Challenge 2015 dataset. Also, they focused on anchor-dependent methods using a compact stacked hourglass network that predicted the centre of the boundary box (but not multi-class instruments) at (37 FPS); this figure is not comparable with those above, since the task is centre-point regression rather than multi-class detection.

Another MIS-related dataset is the Cholec80 dataset [60], which includes phase and tool presence annotations for 80 videos of cholecystectomy. Vardazaryan *et al.* [79] proposed preserving spatial information with a fully convolutional neural network. It predicts instrument presence, and an analysis of the activation maps gives the instrument location. They used a subset of the Cholec80 dataset, selecting images with one instrument per frame since the analysis does not allow multi-instance detection. In 2020, Shi *et al.* [80] at Shandong University took 4011 frames from the Cholec80 dataset and added spatial annotations on the tips of the tools for multi-instance detection. They proposed a two-stage detector, an attention-guided convolutional neural network with coarse and refined modules, to achieve high infer-

ence time (55.5 FPS) and mAP (91.65 %). Cholec80-location subset was also used on a one-stage detector by Yang *et al.* [81], adding modifications to the backbone and neck of the architecture. In the backbone, they used a GhostNet architecture and cross-stage partial connections to increase inference time and enhance the learning process. In the neck of the detector, a U-Net and spatial pyramid pooling are used to address the multi-scale problem. This work reported an mAP of 91.6 % and an inference time of 38.5 FPS. However, there is no free access to this Cholec80-location subset for a fair comparison in the tool detection and classification tasks, limiting the usability and reproducibility of the techniques explored in these works, and preventing these methods from being included in the controlled comparison presented in Chapter 3. Spatial annotations for a subset of the collection have since been released as Cholec80-Boxes [82]. In 2022, Kondo, S. [83] explored the use of a transformer for tool presence without location.

Summary. *Despite substantial progress in surgical tool detection, existing approaches remain limited in their ability to jointly satisfy the representational and operational requirements of real-world MIP. Among the methods reviewed, none reports both high detection accuracy and real-time inference speed, and the figures available are not directly comparable, having been obtained on different datasets and under different evaluation protocols. Furthermore, detection models are often designed around anchor-based assumptions or task-specific heuristics that hinder generalisation and compromise the stability of semantic representations under dynamic surgical conditions. As a result, there remains a gap in detection methods that explicitly address multi-scale variability, discriminative representation learning, and computational efficiency within a unified framework, while producing semantic anchors suitable for downstream temporal tracking and spatial reasoning. Addressing this gap is essential for enabling reliable, real-time surgical perception systems that can operate consistently across procedures, patients, and imaging setups.*

Tissue Identification in MIP

In contrast to surgical instruments, tissues are deformable and visually heterogeneous [52]. As a result, tissue identification is most commonly formulated as a semantic segmentation problem, in which each pixel is assigned an anatomical class, rather than as an instance-level detection or counting task [84]. Obtaining reliable ground truth for anatomical structures is considerably more challenging than for surgical instruments (Figure 2.3), as tissue boundaries are often ambiguous [70], class imbalance is severe, and organs are frequently partially occluded and subject

to substantial inter-patient and intra-procedural variability [27]. Consequently, only a limited number of publicly available datasets provide pixel-level annotations of anatomy in laparoscopic surgery, and these are typically restricted to specific procedures, organs, or imaging conditions [70], [71].

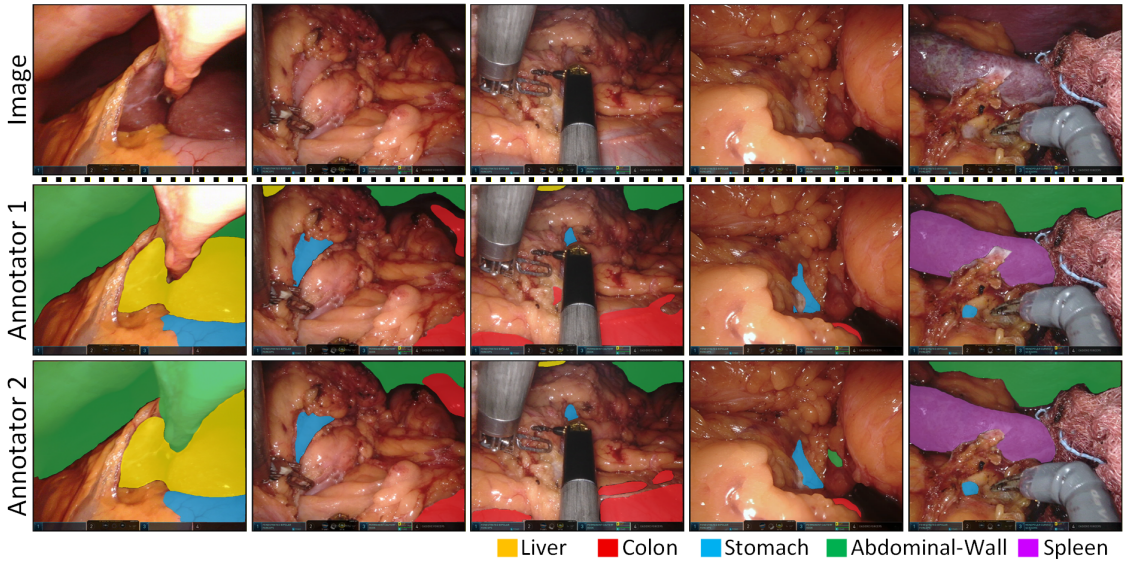


Figure 2.3: **Tissue segmentation**. Example images and segmentation masks from the Dresden dataset [70]. Each image was annotated by two different expert surgeons, illustrating the inherent ambiguity of pixel-level anatomical labelling in surgical scenes.

Existing supervised segmentation approaches can achieve reasonable performance within the confines of individual datasets; however, their ability to generalise across procedures, viewpoints, and degrees of tissue exposure remains limited [85]–[88]. To mitigate the limitations of fully supervised semantic segmentation, weakly supervised and semi-supervised approaches have been explored to reduce annotation requirements and improve the classification of under-represented classes [89]. While current tissue identification methods can alleviate data scarcity and improve robustness and temporal consistency in specific settings, they largely retain pixel-level semantic maps as their core representation [90]. As a result, they continue to face challenges in capturing data for the development and evaluation of approaches that further push the use of these representations into downstream tasks. Recent works, therefore, increasingly shift their motivation towards the exploration of representations that emphasise geometric structure, motion, and temporal consistency over explicit anatomical labelling [51], [52].

Summary. *Beyond data availability, semantic segmentation introduces representational and computational constraints that are misaligned with the perceptual ob-*

jectives considered in this thesis. Pixel-level semantic maps are memory-intensive and computationally demanding, and they do not naturally encode how tissue regions deform, move, or interact with surgical instruments over time. As a result, the trade-off between segmentation accuracy and representational complexity limits their suitability as a foundation for downstream perceptual capabilities in this thesis, including temporal tracking, metric spatial reasoning, and three-dimensional reconstruction. These observations motivate a shift in focus from direct semantic identification of tissue toward modelling its spatial and temporal behaviour. Rather than pursuing increasingly fine-grained pixel-level classification, this thesis emphasises the construction of temporally consistent and geometrically grounded representations that track the motion of points on the tissue surface and recover the metric structure of the scene as it deforms, without modelling the mechanics of that deformation or the interaction between instrument and tissue. Such representations provide a more natural basis for integration across time and viewpoints and align more closely with the requirements of higher-level surgical perception and autonomy. A detailed discussion of temporal and geometric scene characterisation is provided in Sections 2.3 and 2.4. Accordingly, tissue identification is included here for completeness within the perceptual framework, but it is not pursued as a primary task in the remainder of this work.

2.3 Temporal Consistency

Temporal consistency refers to augmenting scene representations with identity persistence over time, addressing the question of how surgical instruments and tissues evolve across successive observations. A representation is temporally consistent, in the sense used throughout this thesis, when it satisfies two properties. The first is identity persistence: an element of the scene retains the same identity, and the same estimated attributes, across the frames in which it appears. The second is path independence: the correspondence estimated between two time steps does not depend on the route taken between them, so that tracking forward and then backward returns the original location, and composing correspondences through intermediate frames agrees with the estimate obtained directly. Inconsistency is accordingly not a general deficiency but the failure of one of these properties, and it is observed as drift, as identity switches between elements, or as estimates that disagree according to the chain of frames from which they were derived. By establishing correspondences between representations across frames, this capability enables the modelling of motion and deformation within the surgical scene. More generally, temporally

consistent representations enable the scene to be interpreted as a continuous process rather than a sequence of independent images, providing a foundation for reasoning about change, causality, and temporal context in surgical workflows.

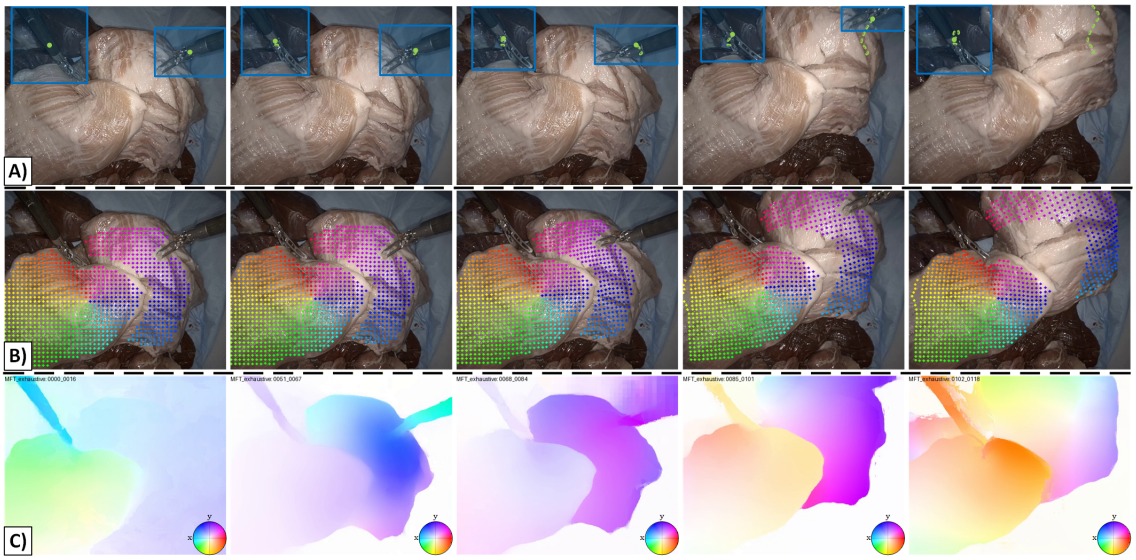


Figure 2.4: **Examples of representation tracking in MIP.** Given a sequence of frames, the form of temporal consistency depends on the representation being tracked. In object detection, this is expressed through the temporal evolution of bounding-box parameters and centroid trajectories (A). At the primitive and segmentation levels, temporal consistency is represented by point trajectories (B), either sparsely defined over selected image or mask locations, or densely estimated through optical flow (C). Raw images in this figure are part of the STIR dataset [91]

Within the proposed framework (Figure 2.1), temporal consistency can be established at multiple representational levels, depending on the abstraction of the instruments and tissues being tracked. Figure 2.4 illustrates how temporal information is manifested across different representations. At the semantic level, object tracking extends detection-based and segmentation-based representations by maintaining the identity of recognised instruments or regions across frames, enabling the estimation of trajectories. Detection and segmentation localise those elements within each frame independently; what tracking adds is the association of those instance localisations across frames, without which a sequence of detections does not constitute a trajectory. At a more primitive level, temporal consistency can be achieved through point tracking, in which local visual primitives extracted from the image are associated over time. This paradigm builds directly on low-level feature representations, and approaches range from frame-to-frame point matching to long-term point trajectory estimation. Together, these tracking formulations provide complementary temporal representations that differ in semantic richness, spatial support, and robustness to deformation.

Temporal consistency tracking is essential for reasoning about motion, deformation, and interaction dynamics [49]. In surgical systems, it enables a range of downstream applications, including force estimation, skill assessment, augmented reality, guided navigation, and increasing levels of autonomy [37], [47], [92]. Nevertheless, temporally consistent representations remain inherently relative: they encode how instruments and tissues move with respect to the image plane or the camera, but not their positions in physical space. Without metric grounding, tracked instruments and tissues cannot be reliably situated in three-dimensional coordinates, limiting their utility for spatial reasoning, measurement, and geometric integration. As such, temporal consistency tracking constitutes a necessary but insufficient step toward unified scene understanding, motivating the introduction of metric spatial reasoning in the subsequent section.

2.3.1 Challenges

Solutions for temporal consistency tracking in surgical scenes must address a range of challenges inherent to MIP, including intermittent image blurriness, variable illumination, complex tissue deformation, occlusion, and homogeneous texture [47]. These challenges hinder the extraction of visual features that remain stable over long image sequences, while anisometric tissue deformation, occlusions and rapid camera or tool movements violate the assumptions of continuous, smooth, and slowly varying motion on which many tracking approaches rely [51]. Furthermore, disentangling camera motion from tissue deformation remains a fundamental challenge, as both contribute simultaneously to observed image dynamics and are difficult to separate using image-based cues alone.

Recent learning-based approaches to point tracking increasingly exploit the idea that dense and temporally consistent correspondences across viewpoints and time translate into three-dimensional localisation and facilitate implicit geometric reasoning [51]. To this end, several methods introduce additional sources of supervision or constraint, such as stereo image pairs, sequential frames, known or estimated camera motion, or three-dimensional reference information, enabling neural networks to learn spatial relationships between visual features. This trend effectively couples correspondence estimation with elements of spatial reasoning, shifting part of the tracking problem beyond purely image-plane associations. It is important to notice that this evolution does not avoid the need for robust two-dimensional correspondences, but rather introduces additional assumptions and requirements related to geometry, sensing, or multi-view consistency. As a result, the incorporation of geometric cues introduces new challenges associated with the availability, reliability,

and integration of spatial information, which must be carefully considered in the context of long-term temporal consistency tracking.

2.3.2 Related Work

Image-Plane Correspondences

Recent learning-based approaches that support a computer vision task hierarchy, which inspire the perception framework presented in this thesis, establish temporal consistency directly in the space of pixels or local visual primitives. Classical pipelines build upon feature detection and description tailored to endoscopic imagery [93]–[95], or correspondence estimation through efficient matching strategies [96]–[100]. Alternatively, new methods learn feature representations and matching functions from data [91], [101]–[103]. These approaches may be trained in a supervised or self-/unsupervised manner, with objectives defined over frame-to-frame consistency, cycle consistency, or photometric alignment. In surgical contexts, unsupervised deep tracking frameworks have been proposed to alleviate the annotation burden by learning correspondences from the regularities that hold between neighbouring frames (photometric consistency, cycle consistency between forward and backward matches, and the smoothness of motion over short intervals) which together supply a supervisory signal in place of annotation [51].

In this family of methods, temporal consistency is still ultimately expressed through image-space associations, but learned representations are used to improve robustness to appearance changes compared to handcrafted descriptors. The central challenge remains the reliable estimation of correspondences under deformation, occlusion, and changing imaging conditions, particularly when temporal horizons extend beyond short frame-to-frame intervals.

Geometry-Aware Tracking

A second family of approaches seeks to stabilise temporal correspondences by introducing geometric constraints or multi-view information. In the surgical literature, this is often discussed in relation to the tight coupling between point tracking and depth estimation [47], [104], where improved correspondence consistency and density can support more stable geometric inference, and conversely, geometric cues can regularise correspondence estimation. Depth estimation is presented in detail in the next section; here, only its use as a supervision signal is discussed. Within this broader family, learning-based methods have explored correspondence and depth estimation with supervision derived from ground-truth 3D information, stereo im-

agery, sequential frames, or camera pose [105]–[109]. These methods effectively couple tracking with implicit spatial reasoning: instead of relying exclusively on explicit feature matching, networks learn representations and similarity functions that are consistent under geometric transformations. This shift has enabled substantial progress in estimating correspondences and geometric structure under constrained assumptions, while introducing new practical requirements related to the availability and reliability of geometric cues.

Camera pose also plays a pivotal role in geometry-aware tracking pipelines and has been recognised as a central challenge for developing methods that extend beyond short temporal windows in surgical settings [110]. Consequently, several approaches incorporate camera motion estimation as part of the learning or optimisation process, drawing upon ideas from Structure from Motion (SfM) and Simultaneous Localisation and Mapping (SLAM) [111]–[113]. Endo-Depth-and-Motion [114] presents a method for camera pose estimation and reconstruction with a self-supervised framework. It uses Monodepth2 [115], photometric tracking, and volumetric fusion to obtain depth information, camera position, and reconstruction, respectively, under the assumption of smooth and slow deformations. A SLAM-based method and the modelling of physiological motion for compensating the patient's respiration during the point-matching stage was presented in [108]; however, it assumes rigid registration. A dataset for validating SLAM methods on endoscopic images was introduced in [116]. The authors also proposed a self-supervised approach "Endo-SfMLearner", which is a combination of residual networks and attention mechanisms to predict camera pose and depth information. Such frameworks highlight an important trend in surgical tracking: temporal consistency can be strengthened by incorporating geometric structure, but doing so also couples the tracking problem to the availability of depth cues, camera motion estimates, or multi-view constraints whose availability depends on the instrumentation, whether the endoscope is stereo or monocular, and whether robot kinematics are recorded.

Rendering-Based Representations

A recent set of approaches further compresses the boundary between tracking and 3D scene modelling by adopting rendering-based representations. NeRF [117] and 3D Gaussian Splatting (3DGS) [118] represent the operative scene in implicit or explicit three-dimensional reconstructions that can synthesise views from novel camera poses, thereby encouraging multi-view consistency across time without supervising correspondences. In surgical contexts, D-NeRF for deformable surgical scenes [119], uses depth maps (preprocessed with STTR-light [120]) as supervision and masks

instruments to facilitate reconstruction of the deforming surgical environment. Similarly, deformable Gaussian splatting has been explored for faster reconstruction and tracking, often using camera pose provided by a dataset and sparse manually labelled correspondences for validation [121]. In parallel, global motion optimisation frameworks have been proposed to enforce consistency across broader temporal windows. Approaches such as Omnimotion [122] formulate motion estimation as a global optimisation problem, and surgical adaptations introduce additional losses or priors to better accommodate specific object categories such as rigid instruments [123]. Collectively, these methods illustrate an emerging direction in which tracking, reconstruction, and view synthesis are treated as interconnected problems, with temporal consistency enforced through shared underlying scene representations and multi-view constraints.

Sensor- and Robot-Assisted Alternatives

Beyond purely vision-based pipelines, several approaches advocate using specialised sensors or markers to obtain three-dimensional information and to stabilise tracking and reconstruction at high frame rates [41], [124], [125]. While such strategies can simplify correspondence estimation by providing direct geometric cues, their integration into minimally invasive surgical environments is often constrained by practical considerations, including workspace limitations, instrument-channel occupancy, and hardware integration restrictions. As a result, most research efforts in surgical tracking have prioritised software-based approaches that rely on standard endoscopic imagery [51].

In robotic minimally invasive surgery, kinematic data can provide partial or complete estimates of camera motion and tool pose, which may be used to inform geometry-aware tracking and to support the disentanglement of camera motion from tissue deformation. These estimates are not, however, expressed in the coordinate system in which the deformation is observed: the kinematic chain reports the pose of the endoscope holder with respect to the robot base, whereas image motion is measured in the camera coordinate system, and the two are related by a rigid transformation that the robot does not know. Recovering it is the Hand-Eye Calibration problem, in which the unknown transform $\mathbf{X}$ between the end-effector and camera coordinate systems is estimated from paired robot and camera motions $\mathbf{A}$ and $\mathbf{B}$ satisfying $\mathbf{AX} = \mathbf{XB}$ [126]. The accuracy of the solution therefore bounds the quality of any disentanglement obtained from kinematics and it must be considered before being compared with motion measured in the image. Robotic platforms such as described in Figure 1.2 may therefore enable multi-modal tracking pipelines that

combine image-plane correspondences with kinematic priors. While the integration of such cues remains underexplored relative to purely image-based tracking, it represents a natural pathway for improving temporal consistency and for supporting downstream metric spatial reasoning and spatial integration within surgical perception systems.

Summary. *The literature reviewed in this section demonstrates substantial progress in temporal consistency tracking and geometric reasoning for surgical perception. Many existing approaches achieve promising results by assuming locally smooth motion, limited deformation, or short temporal windows, conditions under which correspondence estimation and geometric inference can be stabilised. However, these assumptions tend to break down in longer surgical sequences, where tissue deformation is non-linear, discontinuous, and strongly influenced by tool–tissue interaction and camera motion. Under such conditions, tracking performance degrades due to drift, occlusion, and the accumulation of correspondence errors, limiting the applicability of current methods for long-term surgical scene understanding.*

Furthermore, the majority of existing approaches operate primarily in the image plane, enforcing temporal consistency through two-dimensional correspondences without explicitly aggregating information across time to support long-term reasoning. While geometry-aware and multi-view methods introduce spatial constraints to regularise correspondence estimation, they often rely on additional assumptions, specialised sensing, or rigid-motion models that are not well suited to deformable surgical environments. As a result, robust long-term three-dimensional tracking of deformable tissue remains an open challenge, particularly in settings where camera motion and tissue deformation are tightly coupled. Addressing this problem requires approaches that can aggregate short-term, reliable image-space correspondences over extended sequences, extrapolate them into consistent three-dimensional representations, and explicitly disentangle camera motion from tissue dynamics.

In this context, rendering-based methods provide a promising alternative representation. Unlike classical SLAM pipelines, which typically rely on analytical and deterministic optimisation, rendering-based approaches offer learnable, data-driven representations that can adapt to complex appearance changes and non-rigid deformation. However, existing rendering-based methods have primarily focused on view synthesis or short-term reconstruction and have not been fully exploited for long-term deformable point tracking in surgical scenarios. In particular, the integration of test-time optimisation strategies, correspondence aggregation, and robotic kinematic information within a unified rendering-based framework remains largely unexplored.

2.4 Metric and Spatial Grounding

Metric and Spatial Grounding introduce geometric structure into scene representations by associating locations with a three-dimensional coordinate system, thereby answering the question of where elements lie outside the image plane. By grounding image observations in metric geometry, this capability enables reasoning beyond relative motion or appearance change, allowing spatial relationships to be expressed in physically meaningful units. In doing so, geometric grounding enables the scene to be analysed in the same spatial domain as surgical actions, leveraging established geometric and physical models.

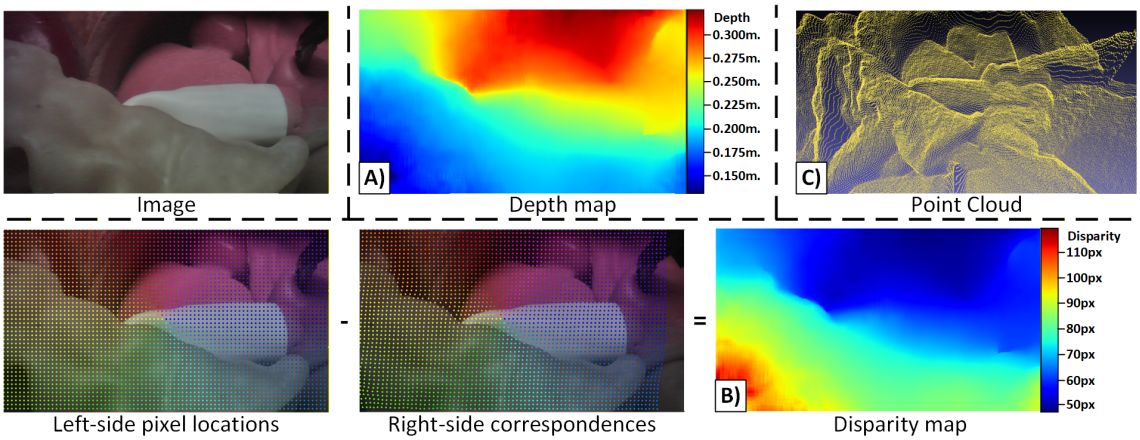


Figure 2.5: **Common metric spatial representations**. (a) Depth maps, (b) disparity maps, and (c) point clouds are commonly used for anchoring representations to a three-dimensional coordinate system. Depth maps encode the distance from the camera to the visible surface for each pixel; disparity maps encode left–right correspondences in binocular endoscopy, which can be triangulated into depth; and point clouds directly associate three-dimensional coordinates with image observations. Raw images and point clouds are part of the DRENDS dataset [127].

Following the proposed perception framework in Figure 2.1, metric spatial reasoning is achieved through depth estimation and camera pose estimation, both of which depend on reliable correspondences [52]. While depth can be estimated from a single frame to recover relative scene structure (i.e., scale-agnostic depth from local primitives), it becomes metrically meaningful only when scale is resolved to obtain metric depth. In practice, scale resolution arises from the interaction between correspondences and camera motion or stereo geometry. Importantly, introducing camera pose augments the scene representation with information not directly observable within a single field of view: it encodes the state of the imaging system and provides the geometric context needed to relate observations across viewpoints. In robotic systems, kinematic information may provide partial or complete pose estimates,

whereas in other scenarios, camera motion must be inferred directly from image data [37]. Through this joint reasoning over depth, scale, and camera pose, scene representations transition from relative image-based descriptions to geometrically grounded spatial structure. Figure 2.5 illustrates common spatial representations used for this purpose.

Metric spatial reasoning is critical in surgical contexts, where accurate spatial relationships underpin navigation, measurement, safety margins, and interaction planning [37], [47]. Knowledge of metric distances, relative depth, and spatial proximity between instruments and anatomy is essential for tasks such as collision avoidance, target localisation, and quantitative assessment of tissue deformation [52]. However, metric information inferred from individual viewpoints or short temporal windows remains inherently local and view-dependent. Without mechanisms to relate camera pose and scene structure consistently across time and viewpoints, metric representations cannot form a persistent or coherent model of the operative environment. As such, while metric spatial reasoning provides the necessary geometric grounding for surgical perception, it must be complemented by mechanisms that align and integrate metric observations over extended temporal and spatial contexts, motivating the spatial integration capability discussed in the subsequent section and reflecting the trends discussed in the previous Section 2.3.

2.4.1 Challenges

Metric spatial reasoning in minimally invasive procedures is fundamentally constrained by the difficulty of acquiring reliable three-dimensional information within the operative environment [52]. Soft tissue undergoes complex, non-rigid deformation driven by respiration and manipulation, all within a confined workspace [27]. Conventional depth acquisition techniques, such as structured light or alignment with preoperative imaging, are often impractical intraoperatively or fail to capture rapid, non-linear motion. The integration of active depth sensors, including laser-based systems such as LiDAR or Time-of-Flight (ToF) sensors, introduces additional challenges related to cost, calibration, sterility, and physical constraints on sensor size and placement [128]. The limited volume of the inflated abdominal cavity further restricts manoeuvrability and sensor integration, making the deployment of specialised hardware for dense 3D sensing difficult in clinical practice.

Consequently, most approaches to metric spatial reasoning in surgery rely on endoscopic imaging. However, inferring depth and scale from endoscopic images is inherently challenging due to the visual characteristics of the surgical scene. Endoscopic imagery is affected by specular reflections, variable illumination, motion blur,

occlusions, and large regions of homogeneous or weakly textured tissue [6]. In addition, tissue deformation violates many of the assumptions underpinning classical geometric methods, such as rigidity or smooth motion [51]. The choice of endoscopic imaging modality further shapes the problem formulation [52]: stereo endoscopes provide two simultaneous viewpoints and enable depth inference via correspondence, thereby providing a partial metric anchor, whereas monocular systems must additionally estimate scale and camera motion over time. Across both settings, solutions must operate under real-time constraints to support intraoperative decision-making.

Overcoming these challenges critically depends on the availability of suitable data for method development and validation [6]. Despite extensive research on depth estimation in computer vision, most publicly available surgical datasets provide depth annotations only for static scenes, rigid phantoms, or highly controlled conditions [52]. In realistic surgical scenarios, obtaining dense, scale-accurate ground-truth depth remains difficult due to continuous tissue deformation, occlusion, and limited access for sensor calibration. This scarcity of ground-truth depth data in dynamic, deformable surgical environments fundamentally limits the evaluation of metric spatial reasoning methods and hinders progress toward robust, temporally consistent 3D understanding in real surgical settings.

2.4.2 Related Work

Data Availability

Several publicly available datasets have supported research in surgical depth estimation, each offering unique contributions while presenting certain limitations. **SERV-CT** [129] comprises 16 stereo-endoscopic image pairs from ex vivo porcine samples, each accompanied by anatomical segmentations derived from cone-beam Computed Tomography (CT) scans. This dataset provides a reliable benchmark due to its high-quality ground truth and detailed anatomical references, albeit with limited diversity and size. The **Hamlyn Laparoscopic/Endoscopic Dataset** [104] provides laparoscopic and endoscopic video sequences, with a subset containing ground-truth depth maps generated using the Libelas stereo matching software on in vivo data. As one of the earliest resources in this field, it remains widely used for benchmarking despite its limited accuracy in depth annotation. **EndoAbs** [130] consists of stereo images and point clouds acquired through a 3D scanner of abdominal organ phantoms. It offers 120 non-sequential, low-resolution frames (640×480), under challenging conditions such as variable lighting and simulated smoke. This makes it valuable for assessing algorithm robustness, although its size and resolution re-

Table 2.1: Comparison of publicly available datasets for depth estimation in minimally invasive procedures.

Dataset	Modality	Data type	Data size	Image resolution	Frame rate	Ground Truth	Camera Motion	Tissue Motion
Hamlyn [104]	Stereo Laparoscopy	In vivo	40,000 images	384×192	unspecified	Libelas stereo matching	True	True
EndoAbs [130]	Stereo Laparoscopy	Phantom	120 images	640×480	3.3 fps	3D scanner	False	False
SCARED [131]	Stereo Laparoscopy	Ex vivo (porcine)	22,950 images	1280×1024	20 fps	Structure from light	True	False
SERV-CT [129]	Stereo Laparoscopy	Ex vivo (porcine)	16 images	720×576	none	CT Scan	True	False
Endo-SLAM [116]	Mono colonoscopy	Ex vivo (porcine)	42,700 images	1350×1080 1280×720 640×480 320×320 256×256	3-35 fps	3D scanners	True	False
SimCol-3D [110]	Mono colonoscopy	Simulation	37,800 images	475×475	20 fps	Virtual Env.	True	False
C3VD [132]	Stereo colonoscopy	Phantom	30,073 images	1350×1080	25 fps	Virtual Env.	True	False

strict broader applicability. **SCARED** [131] dataset provides nine samples, each comprising 4 to 5 short stereo frame sequences. The first frame of each sequence provides high-precision depth information through a structured light approach. On subsequent frames, recorded camera movements and fixed scenarios are leveraged to transform 3D information into their corresponding camera views. This design enables multi-view consistency studies. **Endo-SLAM** [116] provides monocular colonoscopic video sequences accompanied by 6-degree-of-freedom camera poses and dense 3D map ground truth from 3D scanned porcine cadaver organs. The dataset supports the development of visual SLAM techniques in static challenging surgical environments, with a primary focus on colonoscopy. **SimCol3D** [110] is a synthetic colonoscopy dataset generated from multiple trajectories within static 3D colon models. It provides perfectly aligned depth maps and camera poses, offering a controlled environment for algorithm testing. **C3VD** [132] introduces stereo video data of realistic high-fidelity phantom models alongside 3D information per-frame acquired from a corresponding virtual environment. The dataset contains sequences along diverse sections of the colon. By combining realistic surgical imaging with precise geometric annotations, C3VD provides a valuable benchmark for both frame-wise and sequence-based approaches to depth estimation.

Summary. *Table 2.1 summarises the publicly available datasets considered in this section. These resources have enabled progress in surgical depth estimation by providing Red, Green, and Blue (RGB) images alongside depth annotations and, in*

several cases, camera pose information across different viewpoints. They aim to provide a reliable framework for assessing the performance of algorithms that estimate depth maps per frame and algorithms that stitch those maps to reconstruct the surgical scenario. However, existing datasets simplify the challenge of depth acquisition in surgical environments by fixing tissue locations and assuming rigid properties. In practice, tissue motion and non-rigid deformations often arise from surgical manipulation and physiological activity. To the best of our knowledge, no publicly available dataset provides dense, scale-accurate depth ground truth under such dynamic deformable conditions. Data of this form are crucial for developing depth estimation methods that remain reliable over time and for validating their applicability in realistic surgical environments.

Metric Depth in MIP

The approaches taken for depth estimation in minimally invasive procedures can be grouped according to the endoscopic modality: monocular or binocular (stereo) imaging [52].

In monocular endoscopy, metric depth estimation is fundamentally ill-posed from a single frame, as depth must be inferred from motion parallax across time and is therefore tightly coupled to camera motion estimation. Classical pipelines build on structure-from-motion (SfM) and visual SLAM, tracking feature correspondences across frames to estimate camera pose and triangulate sparse structure, which may subsequently be densified through multi-view fusion [56], [111], [113]. More recent monocular approaches increasingly replace explicit correspondence pipelines with learned depth and pose estimation, commonly trained using photometric consistency and self-supervision, and then treat the predicted depth as pseudo-RGBD to enable tracking and fusion. Endo-Depth-and-Motion [104] is a representative example, combining self-supervised depth prediction with photometric tracking and volumetric fusion to estimate camera motion and dense reconstruction from monocular endoscopic videos. Overall, the dominant trend is toward hybrid formulations in which learned depth/pose improves robustness under weak texture and illumination variation, while geometric constraints (e.g., reprojection residuals, keyframe selection, and fusion) provide stabilising structure.

In binocular (stereo) endoscopy, depth can be recovered per frame from disparity between synchronised left–right views, providing an immediate geometric anchor that is absent in monocular settings [129], [131]. Traditional pipelines therefore focus on stereo correspondence (epipolar constraints, typically horizontal after rectification), followed by triangulation, with filtering and (optionally) temporal fusion

to reduce noise and address low-texture or specular regions. Current trends mirror advances in general stereo vision: deep stereo networks learn matching costs and regularisation, improving robustness under challenging appearance conditions, and are increasingly combined with SLAM or fusion back-ends to produce temporally coherent reconstructions over short sequences [107], [120]. A key practical limitation remains that correspondence degrades in homogeneous/specular tissue regions and under occlusions, so even stereo pipelines often require carefully engineered cost aggregation/regularisation and robust filtering to remain stable intraoperatively.

Summary. *While endoscopy remains the primary data source for extracting three-dimensional information in MIP, the scarcity of scale-accurate depth annotations and the prevalence of frame-wise or scale-agnostic formulations continue to constrain the development of methods in realistic deformable scenarios and the capability for metric spatial reasoning. Moreover, progress across the state of the art is often difficult to compare directly, as many works rely on self-constructed datasets and non-standard evaluation protocols, limiting reproducibility and making it unclear how well methods generalise beyond their original settings.*

Chapter 3

Surgical Tool Representations

Within the visual perceptual framework introduced in Chapter 2, surgical tool identification constitutes the first level of representation studied in this thesis. At this level, the goal is to establish stable semantic anchors for the objects in the surgical scene through detection, providing a structured description of the elements and their locations in the image space. In this context, current approaches remain challenged by key characteristics of intraoperative data and operational constraints in MIP. Significant scale variation, strong class imbalance, and high visual similarity among instruments complicate reliable multi-class discrimination, while continuous camera motion and viewpoint changes further affect localisation stability. In parallel, real-time performance requirements impose strict computational constraints, often forcing trade-offs between accuracy, robustness to scale, and inference speed. At the model level, the challenge is to hold local detail and global context at once, since the mechanisms that widen the receptive field trade away the fine detail that separates visually similar instruments or add computation the real-time budget cannot absorb. Additionally, anchor-based detectors need special considerations since they regress boxes against scale and aspect ratio priors set outside the surgical domain.

This chapter addresses these limitations by focusing on the construction of discriminative and structured object representations for surgical tools. Building on transformer-based architectures, the proposed approach learns compact object-level embeddings that summarise salient visual features across multiple scales while maintaining real-time inference. This means that in addition to bounding-box and class prediction, the model produces token-based representations whose embedding space

This chapter is based on the following work: Loza-Galindo, G., Valdastri, P., Ali, S. *Real-time surgical tool detection with multi-scale positional encoding and contrastive learning.* Healthcare Technology Letters, 11:48-58, 2024. DOI: 10.1049/htl2.12060 **Best paper award at the AE-CAI 2023 workshop**

is explicitly supervised, so that representations of the same class are clustered together and those of different classes are separated; the effect of this supervision on the embedding space and on detection accuracy is examined in Section 3.4.

3.1 Problem definition

In this thesis, the problem of surgical tool identification is defined as the construction of object-level representations that jointly encode semantic identity and spatial localisation from a single endoscopic image. Let $I \in \mathbb{R}^{H \times W \times 3}$ denote an RGB image acquired during a minimally invasive procedure. Unlike conventional detection formulations that focus primarily on classification and localisation accuracy, the objective in this chapter is to learn object representations that are both discriminative and structured. In particular, representations corresponding to tools of the same class should be consistent and clustered in an embedding space, while representations of different classes should be well separated. At the same time, these representations must remain robust to scale variation and appearance changes induced by viewpoint, illumination, and partial occlusion, and be computable under real-time constraints.

Formally, the problem consists of learning a mapping f_{obj} that takes as input image evidence and produces a set of object hypotheses $\hat{V} = \{\hat{v}_j\}$, where each hypothesis $\hat{v}_j$ is defined by a triplet (o_j, c_j, b_j). Here, b_j denotes the estimated image-plane bounding box, c_j the associated class label, and $o_j \in \mathbb{R}^d$ an object-level embedding that encodes the visual features of the detected tool. The number of object hypotheses is fixed by the model design, while the effective number of valid detections is determined through a matching process with the ground truth annotations.

Existing detection approaches often learn implicit object representations without explicitly constraining their structure or stability, limiting their suitability as semantic anchors for downstream perceptual tasks. This chapter reformulates surgical tool detection as a joint detection and representation learning problem, in which object-level embeddings are explicitly supervised alongside bounding box regression and classification. Temporal association and metric grounding are beyond the scope of this formulation and are addressed in subsequent chapters. The resulting object representations constitute the first level of structured scene description in the visual perceptual framework, providing a semantic foundation for temporal consistency, spatial reasoning, and integration at later stages.

3.2 Method

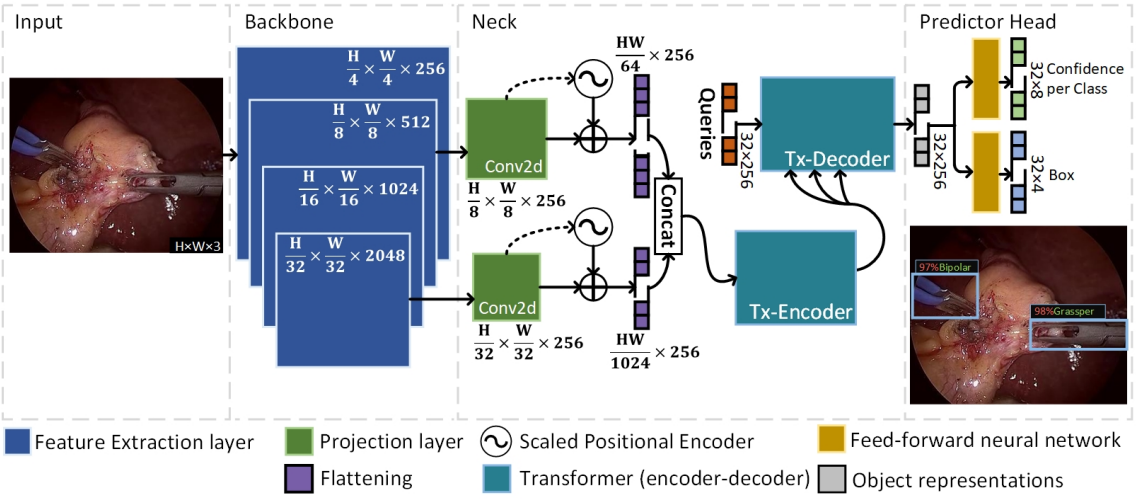


Figure 3.1: **DTX network architecture**. Our proposed DTX architecture uses a Res2Net [133] to extract feature maps at two different scales and forms a dense feature embedding by adding projection layers (Conv2d) with the same number of channels. Then, our network, inspired by DETR [67], uses scaled positional encoders to align features from different layers within a common framework. Finally, the transformer decodes object representations that are fed into two separate feed-forward neural networks for class and boundary-box prediction.

In this work, we propose a new setup for the architecture and training of a multi-scale transformer-based detector (Figure 3.1) that incorporates Res2Net network [133] as a backbone and extracts multi-scale feature maps (from two resolutions), addressing the limitation of small receptive fields and enhancing overall model robustness against scale changes of objects in the images. The extracted features from the backbone are then projected into a common feature space, which enables global multi-scale feature analysis in the Transformer encoder (Tx-encoder). Subsequently, the Transformer decoder (Tx-decoder) generates a set of object representations, which are ultimately processed by two feed-forward neural networks that predict the objects' class and location. In addition to the Hungarian loss used by [67], we proposed a Contrastive loss (CL) for model training. CL loss leverages the Tx-decoder's output to encourage consistency and separability in the generated object representations. This subsection provides a detailed description of the final network architecture and the combined loss functions used in this work.

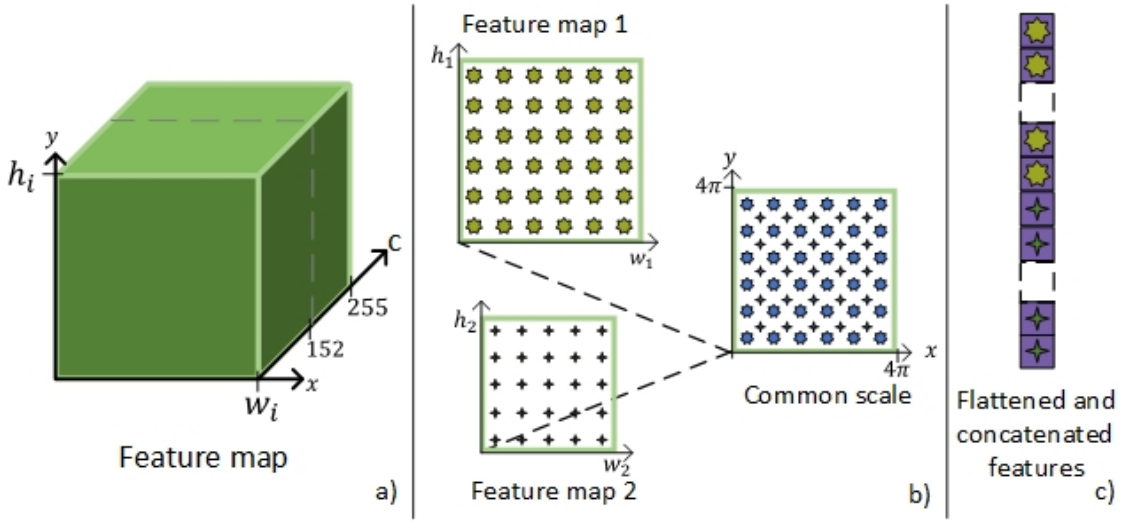


Figure 3.2: **3D Positional Encoder.** a.1) Embedded feature map structure after the projection layer. The position x and y of each feature are encoded in the first and second halves of the channels, respectively. a.2) The positions x, y of the feature maps are scaled by Eq.(3.1) so the transformer can be aware of the location of the features under a common framework. a.3) Input of the transformer: Flattened and concatenated features after the positional encoder.

3.2.1 Architecture details

Similar to the recent DETR network [67], after extracting feature maps from the backbone network, we use a transformer to learn reliable feature representations via self-attention mechanisms. However, our dense transformer (DTX) introduces multi-scale feature connections between the backbone and the Tx-encoder, through additional projection layers and concatenation operations that bring feature maps of different resolutions into a common embedding, and a multi-scale positional encoder (Figure 3.2) that references those features to a single coordinate system in place of the single-scale encoding of DETR. The projection layers (1×1 convolutional layers) reduce the channel dimension to 256, so there is a common feature space between scales (see Figure 3.2.a). It is then the positions (x_j, y_j) of the features from different scales normalised such that they are referenced to a common coordinate system (x, y) despite coming from different resolutions (see Figure 3.2.b). The encoded position for each channel c_k, with k representing the index of the feature channel is given by the sine positional encoder, Eq. (3.1), where width and height are represented as w_s and h_s, respectively, at the scale s of the different feature maps.

$$pos(x_j, y_j, c_k) = \begin{cases} \frac{4\pi}{w_s} x_j & k \in [0, 152] \\ \frac{4\pi}{h_s} y_j & k \in [153, 255] \end{cases} \tag{3.1}$$

Then the embeddings after positional encoding are flatted to a shape $(h_s w_s, 256)$, and these are concatenated along the first axes such as $(\sum_{s=1}^{s=2} (h_s w_s), 256)$ is the final input shape to the transformer encoder (Tx-Encoder) (see Figure 3.2.c). Within the Tx-encoder, the multi-head self-attention modules focus the attention on the features from different scales that are more relevant to the final prediction. In this way, we leverage the transformer for performing attention to both local and global features. The Tx-decoder takes a matrix of zeros as the query to initialise the decoding process. This query shapes the final output by assuming the maximum number of objects in the image and encrypting each object representation in 256 values. The number of queries is therefore treated as a hyperparameter and bounded, and is fixed at 32 by the ablation reported in **Appendix A Table 1**. It is not reduced further to the largest number of instruments annotated in a single frame, because the one-to-one Hungarian matching requires a surplus of queries that can be assigned to the no-object class. Finally, two feed-forward neural networks make the final prediction. A one-layer perceptron with a softmax activation function processes each object representation for its classification, giving the highest probability to the detected object (or not-object class). For the prediction of the boundary box of each object representation, a multi-layer network (3 layers) infers the coordinates of the box (centre x, y, w, and h).

The baseline extracts features at a single scale with ResNet50. Additionally, we tested Res2Net50 [133] defined as a multi-scale backbone (MS) by the authors. Throughout this chapter DTX denotes the transformer using ResNet backbone and MS denotes the use of Res2Net.

3.2.2 Loss functions

We introduce a contrastive loss function in addition to the loss function implemented in DETR [67]. We have a similar matching stage, but unlike DETR, we exploit the matching solution to incorporate the contrastive loss and jointly optimise it with the Hungarian loss.

Matching stage For each image, a $\hat{V}$ set is formed by the predictions of the model and a V set is formed by padding the objects in the Ground Truth (GT) such that both sets have the same number of elements. Each element v_i in V contain (c_i, b_i) where c_i is the class associated with the boundary box b_i and the padded

elements have a c_i value of no-object class ($\varnothing$). Similarly, the element $\hat{v}_j$ in $\hat{V}$ contain $(\hat{o}_j, \hat{c}_j, \hat{b}_j)$ for the object representation, class and boundary box predicted by the model. All the elements in one set are connected to the elements in the second set to form the graph $\mathcal{G}$, thus forming a bipartite graph (Figure 3.3).

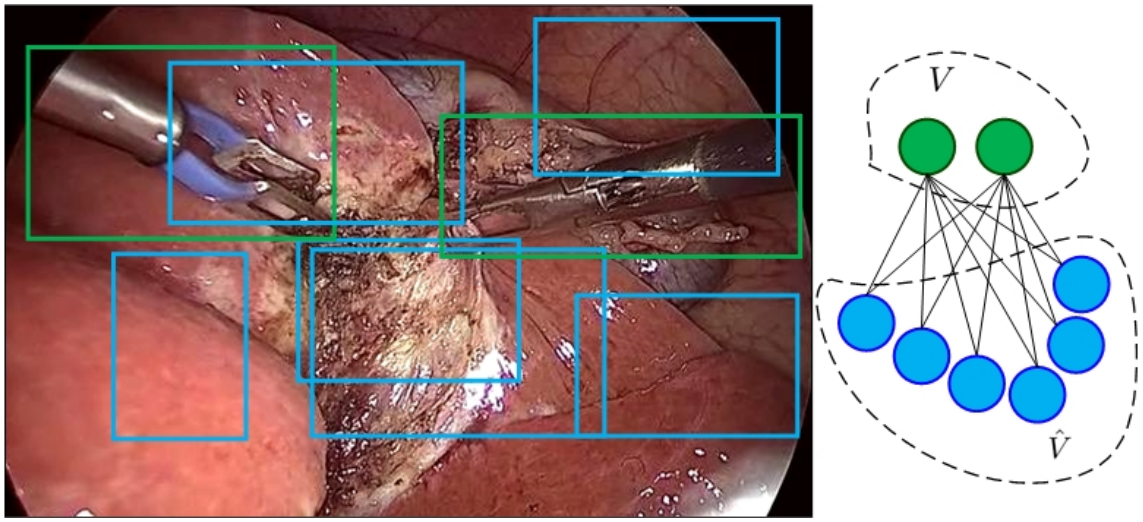


Figure 3.3: **Bipartite graph.** GT boxes in green form the set of vertices V, and predicted boxes in blue form the subset $\hat{V}$. Initially, all elements between sets are connected.

The comparison between the boundary boxes in the GT and the predictions is given by the box loss in Eq. (3.2), where a weighted sum of the L_1-norm and the Generalised Intersection Over Union (GIoU) is used. The matching costs (mc) of a connection (edge) in $\mathcal{G}$ is given by Eq. (3.3), where $\hat{b}_j$ and b_i are boundary boxes (predicted and GT), $\hat{p}_j(c_i)$ is the predicted probability of class i (the GT class) for the predicted box j, and $\mathcal{L}_{box}$ the box loss function.

$$\mathcal{L}_{box}(\hat{b}_j, b_i) = \lambda_1(L_1(\hat{b}_j, b_i)) + \lambda_2(GIoU(\hat{b}_j, b_i)) \tag{3.2}$$

$$mc_{ji} = \mathcal{L}_{box}(\hat{b}_j, b_i) + \lambda_3(1 - \hat{p}_j(c_i)) \tag{3.3}$$

The costs matrix **CM** is then calculated for all samples at indexes i and j by finding the matching cost between the elements of the prediction and the GT. Later, the Hungarian algorithm is used to find unique correspondences between the elements of the sets such that the sum of the matching costs of those correspondences is the minimum. It does that by finding the permutation of the rows in **CM** that minimises the trace of the matrix, so, in the found permutation σ, $\sigma(i)$ is the index j of the matched prediction to the element i in the GT.

Hungarian loss The Hungarian loss function [67] is then applied as shown in Eq. (3.4), which is a weighted combination of the cross-entropy loss and the defined box loss function.

$$\mathcal{L}_H(V, \hat{V}) = \sum_i -\lambda_3 \ln(\hat{p}_{\sigma(i)}(c_i))c_i + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathrm{L}_{box}(\hat{b}_{\sigma(i)}, b_i) \tag{3.4}$$

Contrastive loss We propose adding a complementary contrastive loss ($\mathcal{L}_{CL}$) to our final loss function, jointly optimised with the Hungarian loss. The use of $\mathcal{L}_{CL}$ helps to structure the embedded space by clustering representations with the same class while separating clusters of different classes. The proposed loss is a variation of the Normalised Temperature-Scaled Cross-Entropy Loss (NT-Xent loss) presented in SimCLR [134]. The main difference is that the proposed CL loss can operate within a supervised paradigm, leveraging the Hungarian algorithm's solution. To do so, we look at the samples k in the batch B that contains (V_k, $\hat{V}_k$, σ_k) for the GT, predictions, and the optimal correspondences between them, that is, the permutation σ_k returned by the Hungarian algorithm in the matching stage, which assigns to every element of the GT set exactly one prediction, and we aim to find all the positive $\mathcal{P}$ and negative $\mathcal{N}$ contrastive pairs for each class c in the batch as presented in algorithm 1. $\mathcal{P}_c$ contains all the pairs of object representations (o, o') such that their classes are equal, and $\mathcal{N}_c$ contains all the pairs such that their classes are different. Figure 3.4 displays the contrastive samples for the grasper class in a small batch of 6 images. Note that $\mathcal{P}_c$ avoids the self-comparison, but when the number of representations related to a given class is equal to 1, the pair (o, o) is added in $\mathcal{P}_c$ to pull apart that representation from the rest of the classes in the batch. Then Eq. (3.5) shows the contrastive loss for each class using $\mathcal{P}_c$ and $\mathcal{N}_c$, which applies cosine similarity *sim* between the object representations.

$$\mathcal{L}_{CL_c}(\mathcal{P}_c, \mathcal{N}_c) = -log \frac{\sum_{(o,o') \in \mathcal{P}_c} exp(sim(o, o'))}{\sum_{(o,o') \in (\mathcal{P}_c \cup \mathcal{N}_c)} exp(sim(oi, o'))} \tag{3.5}$$

The total contrastive loss is the average of all the contrastive losses per class in a given batch B with nc classes and size bs. Thus, the final loss $\mathcal{L}$, which is an equally weighted sum of the Hungarian loss and the contrastive loss, can be represented as:

$$\mathcal{L}(B) = \frac{\lambda_4}{bs} \sum_{k=0}^{bs} \mathcal{L}_H(V_k, \hat{V}_k) + \frac{\lambda_4}{nc} \sum_{c=0}^{nc} \mathcal{L}_{CL_c}(\mathcal{P}_c, \mathcal{N}_c). \tag{3.6}$$

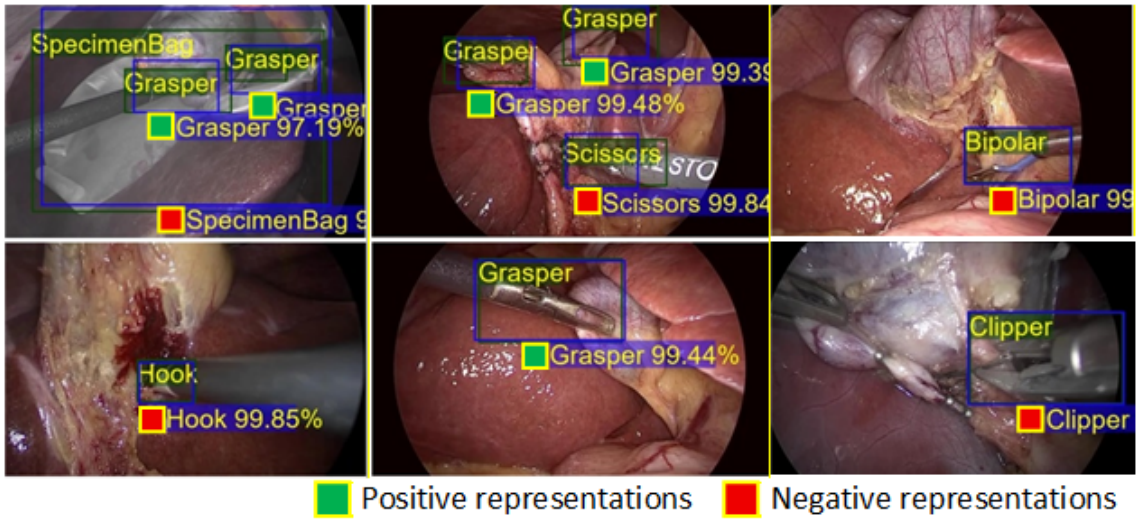


Figure 3.4: **Contrastive object representations for the grasper class.** The object representations of the boxes that were correctly predicted (in terms of location and class) are used to find contrastive samples for each class. In this case, the grasper class is analysed, green and red markers highlight positive and negative samples respectively.

3.3 Experiments and results

3.3.1 Dataset

We evaluate our architecture on the publicly available m2cai16-tool-location dataset [73] containing 2,532 labelled frames from 15 videos of cholecystectomy procedures performed at the University Hospital of Strasbourg in France. To make our method comparable and reproducible, we have used the same split proposed in the original paper [73]. The final experimental dataset comprises 1405 images for training, 843 images for validation, and 563 images for testing (held-out set). As Sahu [135] pointed out, this dataset poses an extra challenge to a solution for the multi-class classification problem since it mirrors the imbalance appearing of the surgical tool during the operation. Therefore, the seven tool classes were considered in the ground truth labels, together with the no-object class ($\varnothing$) defined in Section 3.2.2, and a discussion on how the implemented solution alleviates this problem is presented in the results section.

A second dataset is used for evaluation only. Cholec80-Boxes [82] provides bounding box annotations over five videos of the Cholec80 collection, recorded at the same centre and covering the same seven tool classes. It comprises 14195 annotated frames and 23708 boxes, an order of magnitude more than the training split used here, and its class distribution is the one observed during the procedure rather

Algorithm 1 Supervised contrastive learning algorithm for multi-class labels.

Require: Batch: B; Classes: C
  $n_c = 0$ ▷ number of classes
  $\mathcal{L} = 0$
  for c $\in C$ **do**
    pos_samples = neg_samples = {}
    for $k \in \{0$ to $len(B)\}$ **do**
      for $i \in \{0$ to $len(V[k])\}$ **do**
        if $V[k].c[i] ==$ c **then**
          pos_samples $\leftarrow \hat{V}[k].o[\sigma[k](i)]$
        else
          neg_samples $\leftarrow \hat{V}[k].o[\sigma[k](i)]$
        end if
      end for
    end for
    $\mathcal{P} = \mathcal{N} = \{\}$ ▷ positive & negative contrastive pairs
    for $o \in pos_samples$ **do**
      for $o' \in neg_samples$ **do** $\mathcal{N} \leftarrow (o, o')$
      end for
      $pos_samples.pop(o)$ ▷ Remove the reference from the list
      for $o' \in pos_samples$ **do** $\mathcal{P} \leftarrow (o, o')$
      end for
      if $\mathcal{P} == \emptyset$ **then** $\mathcal{P} \leftarrow (o, o)$
      end if
    end for
    $n_c = n_c + 1$
    $\mathcal{L} = \mathcal{L} + Eq5(\mathcal{P}, \mathcal{N})$
  end for
  $\mathcal{L}_{contrastive} = \mathcal{L}/n_c$

than a curated balance: the grasper and the hook account for 83% of the boxes and the scissors and the clipper for 1.2% each. No model is trained or tuned on it. It is used in Section 3.3.4 to measure how the configurations of Table 3.1 transfer to patients and annotations they have not seen.

3.3.2 Experimental setup

Data augmentation. All images were resized to 320×320 pixels. Six different geometric transformations were selected for data augmentation. During training, the transformations were randomly applied with a 33% probability each.

Model configuration. The optimal hyperparameters for our model are reported in this section. However, a hyperparameter search grid is presented in the ablation study. The building blocks in the Res2Net50 architecture (the used backbone) were

configured to split the feature maps into four sets of 26 channels each. In the neck of our architecture (see Fig. 3.1), the feature maps that go through the projection layers were taken from layers 2 and 4 of the backbone. The number of queries that initialise the decoder process in the transformer was set to 32, and the number of layers in the encoder and decoder of the transformer to 6.

Training setup. We build our model leveraging part of HuggingFace's Transformers repository [136] and making the pertinent changes to match the model's description presented in Section 3.2. During training, an AdamW optimiser with a step learning rate scheduler was added. The scheduler tracked and modified the learning rate from $1.0e^{-04}$ to $1.0e^{-06}$, with a factor of 0.5 at every 40 epochs. In addition, a stopping criterion tracking the validation loss was included in the experiment. It had a patience of 50 epochs and considered a minimum delta of $1.0e^{-0.5}$. We run all our code in a setting with multiple CPU processors provided by the Research Computing Team at the University of Leeds in their High-Performance Computing (HPC) facilities. The requested nodes provided 48GB system memory and an NVIDIA V100 32GB graphic card.

Evaluation metrics. We present and compare the performance of our model based on two widely used metrics called mean average precision (mAP) for object detection. For this metric, a threshold value is used to determine if detection is considered a true positive or a false positive based on the Intersection Over Union (IoU) value ranging from [0.5 : 0.05 : 0.95] for overall mAP and at specific IoUs, e.g., [0.5] and [0.75]. The second metric reported is the inference time in frames per second (FPS).

Ablation protocol. Because average precision is computed once over the whole split rather than per image (as in the standard COCO protocol), the sensitivity of a reported value to the particular frames that were annotated is estimated by resampling. 2000 bootstrap replicates are drawn, each of the same size as the split and sampled with replacement, taking the whole image as the resampling unit so that objects sharing a frame are resampled together. Every configuration is evaluated on the same replicates and each difference between two configurations is taken within a replicate, so that the difficulty of the draw cancels, and the standard deviation is taken over the replicates.

3.3.3 Comparison with SOTA and baseline methods

In this Section, we provide a comparison with state-of-the-art methods used for detection tasks on the m2cai16-tool-location dataset. Alongside this, we also present quantitative results on the baseline model and provide results for different architec-

Table 3.1: **Quantitative results.** Comparison of state-of-the-art surgical tool detection methods, anchor-free methods, and our proposed dense transformer (DTX) with and without multi-scale and contrastive loss inclusions.

Model	$\mathbf{mAP}_{[50:95]}$	$\mathbf{mAP}_{[50]}$	$\mathbf{mAP}_{[75]}$	FPS	Backbone	Input size
SOTA comparison						
Fast R-CNN [73]	NA	0.631	NA	5	VGG-16	NA
Fast R-CNN [74]	NA	0.696	NA	15*	ResNet101	NA
YOLO [75]	NA	0.722	NA	48	DarkNet19*	448×448
DSSD [76]	NA	0.912	NA	15*	ResNet101	320×320
F. R-CNN+SSL[137]	0.468	0.902	0.462	15*	ResNet50-FPN	NA
Anchor free methods						
FCOS[66]	—	0.900	—	12	ResNet50	450×450
DETR[67] (baseline)	0.520	0.886	0.524	**36**	ResNet50	320×320
Our proposed approaches						
DTX	0.536	0.926	0.557	35	ResNet50	320×320
DTX + MS	0.543	0.939	0.561	32	Res2Net50	320×320
DTX + MS + CL	**0.545**	**0.945**	**0.572**	32	Res2Net50	320×320

DTX, Dense Transformer (our method); Contrastive Loss, CL;MS, Multi-scale backbone (Res2Net)
NA, not available; * This value was not officially reported by the author.

tural changes that have been proposed.

Quantitative results. Tables 3.1 and 3.2 present comparisons of the SOTA methods for supervised surgical tool detection, anchor-free methods in the literature, and our proposals for overall mAP and AP for each class category, respectively. From Table 3.1, it is evident that our proposed approaches outperformed both the SOTA methods and other anchor-free methods. For example, our final model (DTX+MS+CL) has $\text{mAP}_{[50]}$ is 4% above the best SOTA method (DSSD), and nearly 7% higher than the baseline DETR. Our experiments also showed an additional improvement at $\text{mAP}_{[75]}$ over the baseline with 0.572 compared to 0.524, which is 9% above. On the FPS, our method achieves 113% faster than the SOTA DSSD method and is only slightly lower than DETR baseline methods (4 FPS). Table 3.2 shows consistent improvement in all class categories compared to the SOTA and the baseline DETR, regardless of the frequency with which each tool class appears in the dataset's images. Common (e.g., the grasper and hook) and rare (e.g., scissors and bipolar) tools are detected with high mAP, which suggests that the model focuses on relevant features from the images for the formation of the object representations associated with each class. Section 3.3.4 shows that this uniformity is a property of the training distribution and does not survive a change of dataset. For example, compared to the most accurate method, DETR, our approach achieves 8%, 7.7%, 11%, 3%, 8.8%, 2.7%, and 5.2%, respectively, for grasper, bipolar, hook, scissors, clipper, irrigator, and specimen bag.

Qualitative results. Figure 3.5 shows predictions from our proposed approach (DTX+MS+CL). The selected samples were the images with very low errors (on the left) and the images with the most significant errors (on the right). It can

Table 3.2: **Quantitative results**. Average precision (AP) comparison per class.

Method	Grasper	Bipolar	Hook	Scissors	Clipper	Irrigator	S. Bag
SOTA comparison							
Fast R-CNN [73]	0.483	0.670	0.784	0.677	0.863	0.175	0.765
Fast R-CNN [74]	0.541	0.695	0.868	0.739	0.842	0.416	0.771
YOLO [75]	0.893	0.324	0.932	0.666	0.903	0.424	0.914
Anchor free methods							
FCOS [66]	0.846	0.927	0.942	0.905	0.903	0.857	0.922
DETR [67] (baseline)	0.826	0.910	0.864	0.915	0.844	0.932	0.911
Our proposed approaches							
DTX (ours)	0.871	0.957	0.933	0.900	0.926	0.942	0.950
DTX + MS	0.891	0.955	0.955	**0.965**	**0.933**	0.921	0.956
DTX + MS + CL	**0.894**	**0.980**	**0.960**	0.945	0.919	**0.957**	**0.959**

DTX, Dense Transformer (our method); Contrastive Loss, CL; MS, Multi-scale backbone (Res2Net)

be observed that for the frames with optimal predictions, the predicted boxes (in blue) completely overlap the ground truth boxes (in green). However, for those with erroneous predictions (in the right), in most cases, either the object was not present (frame incorrectly labelled) or the object was incorrectly classified due to the fact that the intrinsic characteristics of the object are not present. In the second case, we can observe that our model makes a good guess by associating the object with a fairly similar tool. Figure 3.6 shows that the object representation space generated by our model in the decoder of the transformer is organised after implementing contrastive learning by maximising the distance between the clusters of the classes and arranging misclassified objects. This adds up to the AP improvement presented in Tables 3.1 and 3.2, strongly suggesting that the error due to the misclassification of objects is considerably alleviated with our approach while boosting the performance. Having solved this problem, future efforts could be focused on developing methods that

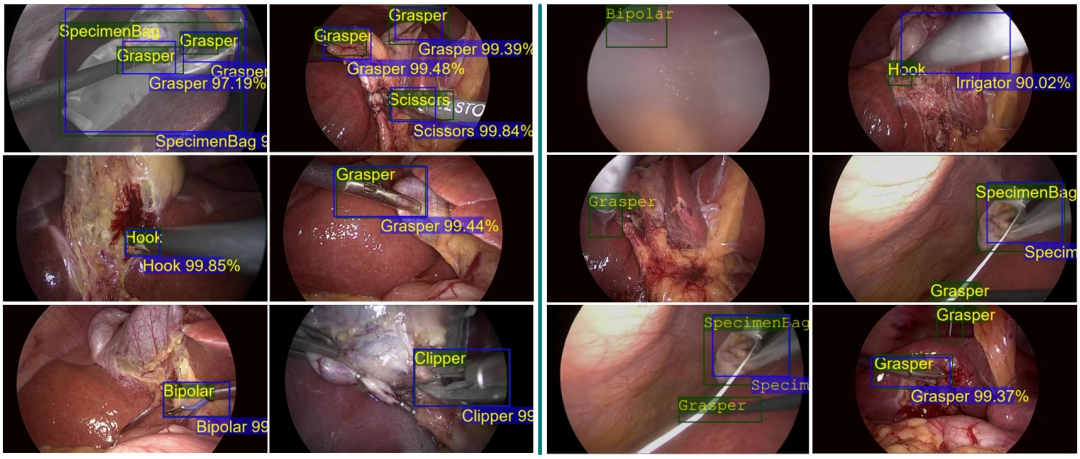


Figure 3.5: **Qualitative results:** Frames taken from the test set with their respective predictions. Predictions with the highest IoU are presented on the left, and predictions with the lowest IoU are presented on the right.

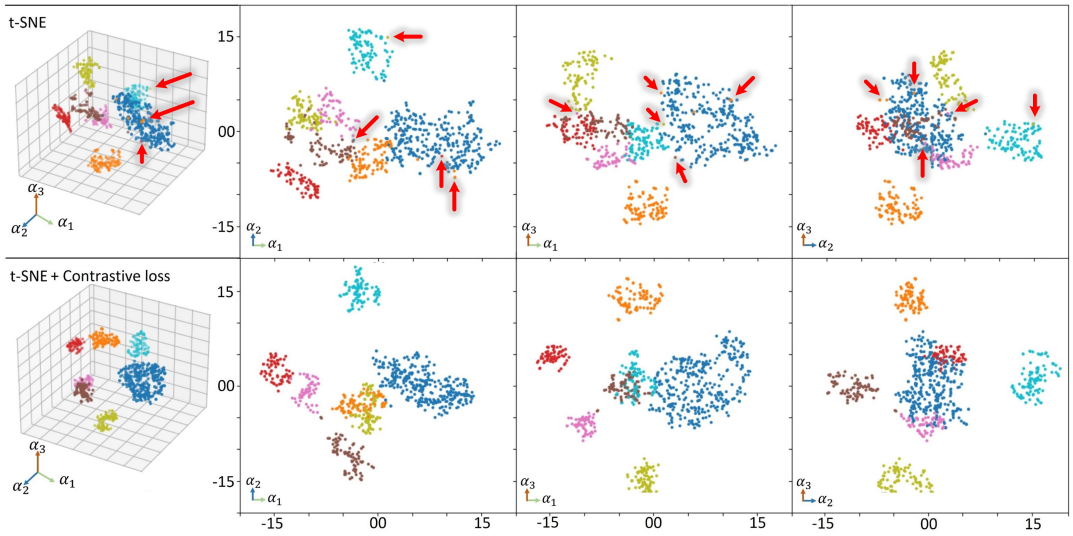


Figure 3.6: **Object representation space.** Dimensional reduction of the object representation space (Tx-decoder's outputs) using t-SNE, each dot in the graphs represents a detected object by our model (DTX+MS). At the top, without contrastive learning (CL), the clusters for each class are barely separated, and some points are mislocated (see red arrows). This distribution is prone to hindering the performance of classifiers. However, at the bottom, we can clearly see how the integration of the CL alleviates this problem. There is a wide separation between clusters, and all the mislocated points were correctly rearranged.

increase the precision of the predicted boundary box so that the value in the IoU is improved. **Appendix A Figure 1** shows the attention maps from the transformer's last layer in the decoder. Since we use feature maps at different scales, these images demonstrate how the relationship between the features in the regions of attention is present at different scales.

Ablation study. The performance of models over the validation set for different network configurations (e.g., scales, feature layers, and feature maps) and combinations of relevant hyper- parameters (e.g., number of queries) is presented in **Appendix A Table 1**. It can be observed that for different numbers of queries ranging from 32 to 100 queries, 32 queries boosted the performance of the model on the $mAP_{[75]}$ by 8.8% compared to using the number of queries proposed by DETR [67]. Our experiments also showed that a combination of four scales and 26 channels is the optimal backbone yielding 6.2% and 3.6% of improvement on the $mAP_{[50]}$ and $mAP_{[75]}$, respectively. The number of layers in the encoder and decoder of the transformer shows that a network with six layers provided the best trade-off between accuracy (0.866) and inference speed (FPS of 36). Finally, the contribution of the multi-scale backbone (MS) and of the contrastive loss (CL) is examined on the test split with the resampling protocol described in Section 3.3, and the paired

Table 3.3: **Ablation significance.** Paired differences between the ablation models on the m2cai16-tool-locations test split ($n = 563$), obtained from 2000 bootstrap replicates of the split. SD is the standard deviation of the difference across the replicates, and Positive Δ the fraction of replicates in which the difference is positive.

Comparison	Δ mAP$_{[50]}$	SD	Positive Δ
DTX vs DETR	+0.039	0.011	1.000
DTX + MS vs DTX	+0.013	0.009	0.931
DTX + MS + CL vs DTX + MS	+0.009	0.008	0.886
DTX + MS + CL vs DTX	+0.022	0.010	0.994
DTX + MS + CL vs DETR	+0.060	0.012	1.000

DTX, Dense Transformer (our method); CL, Contrastive Loss; MS, Multi-scale backbone (Res2Net).

differences between adjacent configurations are reported in Table 3.3.

The full model improves on DTX by 0.022 mAP$_{[50]}$, with the difference positive in 99.4% of the replicates, so the gain of the complete architecture is not carried by the dense transformer alone. Taken one at a time, however, neither step can be certified on this split, since at 0.013 and 0.009 each is of the order of one standard deviation. Each nevertheless remains positive in 93.1% and 88.6% of the replicates, against the 50% expected if the component made no difference, so the evidence is consistent in direction rather than absent, and it is the size of the published test split that bounds what can be claimed from it.

3.3.4 Cross-dataset evaluation

The four configurations of Table 3.1 were evaluated on Cholec80-Boxes without retraining, fine-tuning or the adjustment of any threshold, so what is measured is how the architecture behaves on patients, instruments and annotations it has not seen. Because parts of the literature treat the m2cai16 videos as drawn from the same pool as Cholec80, the two collections were first compared frame by frame on contrast-normalised thumbnails: the closest cross-dataset pair lies 27 times further apart than a re-encoded copy of the same frame, so no frame is shared. The overall results are reported in Table 3.4 and the per-class results in Table 3.5, in the terms of Tables 3.1 and 3.2.

A drop is expected when a detector trained on one collection is evaluated on another, since the appearance of the scene and the annotation protocol change, and every configuration loses about half of its accuracy, the best of them falling from 0.939 to 0.594. PArt of it is an artefact of the comparison. The boxes of Cholec80-Boxes are drawn tighter than those of m2cai16-tool-locations, and rescaling every

Table 3.4: **Cross-dataset evaluation.** The four configurations of Table 3.1 evaluated on Cholec80-Boxes (14195 frames, 23708 boxes) without retraining or tuning of any kind. The last column repeats the $mAP_{[50]}$ of Table 3.1 for reference.

Model	$mAP_{[50:95]}$	$mAP_{[50]}$	$mAP_{[75]}$	m2cai16 $mAP_{[50]}$
DETR (baseline)	0.189	0.454	0.130	0.886
DTX	0.214	0.547	0.131	0.926
DTX + MS	**0.246**	**0.594**	**0.162**	0.939
DTX + MS + CL	0.224	0.557	0.141	0.945

DTX, Dense Transformer (our method); CL, Contrastive Loss;
MS, Multi-scale backbone (Res2Net).

Table 3.5: **Cross-dataset evaluation.** Average precision ($AP_{[50]}$) per class on Cholec80-Boxes, in the terms of Table 3.2. Instances per class in this set: grasper 10399, hook 9278, irrigator 1396, specimen bag 1091, bipolar 982, clipper 285, scissors 277.

Model	Grasper	Bipolar	Hook	Scissors	Clipper	Irrigator	S. Bag
DETR (baseline)	0.541	0.655	0.555	0.161	0.574	0.232	0.459
DTX	**0.638**	0.706	0.731	0.149	0.598	**0.409**	0.601
DTX + MS	0.637	0.718	0.772	**0.357**	**0.687**	0.330	**0.658**
DTX + MS + CL	0.535	**0.770**	**0.781**	0.291	0.681	0.206	0.637

DTX, Dense Transformer (our method); CL, Contrastive Loss; MS, Multi-scale backbone (Res2Net).

prediction about its own centre recovers at most 0.020 of the gap at an optimal factor of 0.90; the median intersection over union of matched pairs is 0.726, so the model localises well what it finds and the gap is carried by the tools it misses or misclassifies.

What survives the change of dataset is the advantage of the proposed architecture over the baseline. Every configuration remains above DETR, and the complete model improves on it by 0.103 $mAP_{[50]}$ here against the 0.060 of Table 3.3, so the benefit claimed in this chapter is wider on unseen data than on the published split. The ordering between the proposed configurations is not preserved. DTX + MS improves on DTX by 0.047 $mAP_{[50]}$, and the improvement is carried by the classes with the fewest instances in this set, 0.208 on the scissors, 0.089 on the clipper and 0.057 on the specimen bag, which together hold 7% of the boxes. Conversely, DTX + MS + CL falls 0.037 below DTX + MS, so the most accurate model on unseen data is DTX + MS. The contrastive term keeps only two of its per-class gains, the bipolar and the hook, while its largest gain in Table 3.2, 0.036 on the irrigator over 84 test instances, reverses to a loss of 0.124 over the 1396 instances of this set.

3.4 Discussion and conclusions

Even though there are works in surgical tool detection in literature, these methods are widely built on anchor-based methods, do not incorporate multi-scale feature embedding for tackling variable tool sizes, and suffer from low speed [73]–[76], [137]. Our approach using a transformer with the incorporation of multi-scale feature selection is not only independent of anchors but also provides improved accuracy and inference time compared to SOTA methods in the literature. Utilising the Res2Net backbone into our proposed dense transformer (DTX) enabled the inclusion of local and global features that can jointly tackle variations in the size of the objects and receptive field constraints. Our experiments showed improvement in almost all the tool categories by a large margin, up to 10.5%, compared to the baseline model (DETR [67]), which is the most consistent across the tool categories compared to any SOTA methods (Table 3.2). Further, we also showed that the incorporation of contrastive loss aids in clustering the representations of a class and separating the clusters of different classes (Figure 3.6). Without contrastive learning the class clusters overlap and several representations fall outside their cluster, whereas with it the clusters are distinct and those points are rearranged (Figure 3.6). The contribution of the loss is therefore one of representation, and the value of object representations that remain consistent within a class lies in the tasks that depend on the identity of an instrument being recovered over time, such as tracking. The cross-dataset evaluation of Section 3.3.4 bounds that contribution. Before the contrastive term the regions that the classes occupy in the representation space are broad and overlapping, so a tool whose appearance falls outside the training distribution still lands inside one of them, and often inside the right one; after it, those regions are tight and the same tool can land in the gap between them. Consistency therefore does not make the detection easier, it makes the discrimination between classes easier, and the two are different tasks. Measuring what the contrastive term does to the representation space under a change of dataset would require a separate investigation, which is not attempted here. The full model, which combines the multi-scale backbone with this loss, improves average precision in every tool category over the baseline detector, from 2.7% for the irrigator to 11% for the hook (Table 3.2). The less accurate predictions of our model are probably due to the fact that there are not enough intrinsic features of the object within those samples, and confusion might happen, for example, misclassification of grasper and clipper (Figure 3.5). Consideration of features from previous frames could alleviate this problem and boost a more accurate prediction.

In conclusion, we proposed a transformer-based surgical tool detection method

introducing a novel multi-scale feature assembly and incorporation of contrastive loss function utilising information from the bipartite graph. The proposed model is anchor-free and has near real-time performance (32 FPS). The contrastive term structures object representations by class, and this thesis contributes to building reliable and consistent representations through that structure. To this extent, we also demonstrated the superiority of our approach compared to several SOTA approaches and other anchor-free methods. The qualitative results also demonstrated the effectiveness of our model, with high-quality predictions even in the challenging scenes. In our future work, we aim to leverage video temporal features to improve tool detection. Bounding box annotations for five videos of the Cholec80 collection have since been released as Cholec80-Boxes [82], covering the same seven instrument categories used here, and the configurations of Table 3.1 were evaluated on them without retraining. Every one of them retains its advantage over the baseline detector, and DTX + MS is the most accurate of them, which places the multi-scale backbone rather than the contrastive term at the centre of what generalises (Section 3.3.4).

Chapter 4

Temporal Consistency: Long-term 2D and 3D tracking

Within the visual-perceptual framework introduced in Chapter 2, temporal consistency is a critical step toward constructing dynamic, spatially coherent representations of the surgical scene. At this stage, the objective is to associate visual information across time, enabling the modelling of motion and deformation beyond static observations. Temporal consistency can be established at multiple representational levels, ranging from associating low-level visual primitives to tracking semantic objects. In this thesis, temporal consistency is investigated at the level of image-space primitives, where point-based representations provide a flexible basis for tracking deformable structures under challenging surgical conditions.

Achieving reliable temporal consistency in MIP remains challenging due to the characteristics of intraoperative data and the dynamic nature of the surgical environment. Intermittent motion blur, variable illumination, occlusion, and homogeneous textures hinder the extraction of stable visual features over long sequences, whereas non-linear tissue deformation and rapid camera or tool motion violate the assumptions of smooth, slow change commonly adopted by state-of-the-art tracking approaches. This chapter addresses these limitations by focusing on the construction of long-term correspondences through the aggregation of reliable short-term two-dimensional correspondences, stereo-based depth estimation and robot-informed camera pose within a single framework. The proposed method employs a test-time optimisation framework that enhances neural radiance field–based scene representations by embedding correspondences directly into the implicit model. It does so

This chapter is based on the following work: Loza, G., Hu, J., Jones, D., Ali, S., Valdastri, P. *Surg-InvNeRF: Invertible NeRF for 3D tracking and reconstruction in surgical vision.* 2025. (Under review). https://www.arxiv.org/pdf/2508.09681. Project page: STORMLAB-InvNeRF.

through joint reasoning across image space, the deformable workspace, and a canonical space, while explicitly separating camera motion from tissue dynamics. This contribution supports stable correspondence aggregation under deformation, rapid motion, and occlusion. Temporal consistency is used here in the sense defined in Chapter 2: a representation is temporally consistent when an element of the scene retains its identity and its estimated attributes across the frames in which it appears, and when the correspondence estimated between two time steps does not depend on the route taken between them. It is this second property, path independence, that the present chapter carries. It is built into the architecture, since the mapping into the canonical space is invertible and a correspondence followed forward and then backward therefore returns the original point exactly, and it is enforced during optimisation by the cross consistency loss, which requires correspondences that reach the same scene point through different frames to coincide in the canonical space. Its effect is measured in the ablation of Table 4.6, where enabling this loss reduces the 2D tracking error from 18.170 to 14.421 pixels and the 3D tracking error from 9.736 to 6.582 millimetres. These temporally consistent representations extend the expressive capacity of local and primitive representations within the visual perceptual framework presented in this thesis and provide evidence for the importance of spatially and metrically grounded representations, as further discussed in subsequent chapters.

4.1 Problem definition

We define the problem of establishing frame-to-frame correspondences for a long sequence of consecutive video frames as follows. Let $\mathbf{I} = \{I_0, I_1, I_2, ..., I_t\}$ denote the sequence of RGB images, where $I_t \in \mathbb{R}^{H,W,3}$ for every time t. Let $\mathbf{p}_t$ represent pixel locations within I_t. The function f_{ST} can be any feature matcher function that identifies reliable short-term correspondences between frames, predicting the location $\mathbf{p}_{t_1 \to t_2}$ in I_{t_2} matching $\mathbf{p}_{t_1}$ in I_{t_1}. Such function is then defined by $\mathbf{p}_{t_1 \to t_2} = f_{ST}(\mathbf{p}_{t_1}, I_{t_1}, I_{t_2})$ with $||t_2 - t_1|| < k$ and where k is the maximum time window between short-term pairs. Given a set Ω of short-term correspondences $(\mathbf{p}_{t_1}, \mathbf{p}_{t_1 \to t_2})$ for every t in a sequence of frames, we look for a long-term correspondence function $f_{LT}(\mathbf{p}_{t_1}, t_1, t_2) = \hat{\mathbf{p}}_{t_1 \to t_2}$ that minimise the error between $\hat{\mathbf{p}}_{t_1 \to t_2}$ and all the possible chains to go from t_1 to t_2 using elements in Ω. A correspondence is short-term when the two frames it relates fall within the window k, so that it can be estimated directly by f_{ST}, and long-term when they do not, in which case no direct estimate is available and the correspondence has to be inferred from chains of short-term pairs.

The two functions are therefore distinguished by their domain rather than by the direction of time: f_{ST} is defined only for pairs closer than k, whereas f_{LT} is defined for any pair of time steps in the sequence.

Omnimotion [122] proposed a rendering-based framework for the parametrisation of f_{LT}. However, this approach assumes an orthographic camera model and an unbounded temporal window for short-term pairs, making it computationally expensive and inefficient. Thus, the current SOTA does not provide reliable correspondences from f_{ST} when $|t_2| >> |t_1|$, and the 3D-rendered positions do not represent actual locations in the workspace. Along these lines, FastOmniTrack [138] improved speed and accuracy by removing the rendering stage, using precomputed scale-agnostic depth, reducing the temporal window size, and introducing a non-linear, invertible function. Despite the improvement, it retains the orthographic camera model and uses non-metric depth, which prevents 3D tracking of points in the scene. In this work, we propose to parametrise f_{LT} using a pinhole camera model in our Invertible NeRF to fully recover the benefits of 3D reconstruction while tracking points in 3D. We also reduce memory requirements by constraining the maximum number of pairs per time step, thereby speeding up the optimisation of the function (i.e., reducing the number of iterations).

4.2 Methodology

This section describes our proposed NeRF-based framework for test-time optimisation that aggregates the 3D projections of short-term 2D pixel correspondences to enable long-term 3D tracking. A description of the data and the optimisation flow for the proposed framework is shown in Figure 4.1. Here, f_{LT} is mainly defined by the combination of three core components: 1) a ray sampler ϕ_{sampl} (red block), 2) a colour and density network $\phi_{c/\sigma}$ (blue block), and 3) an invertible deforming network ϕ_{def} (green block). Appendix B.1 provides further details about the reasoning on the architecture design.

During the optimisation, the data from pairs of frames (I_{t_1} and I_{t_2}) with short-term pixel correspondences in Ω is passed. It includes frames' time steps (t_1, t_2), 2D pixel locations $(\mathbf{p}_{t_1}, \mathbf{p}_{t_2})$, pixel colours $(\mathbf{c}_{t_1}, \mathbf{c}_{t_2})$, extrinsic camera parameters $(\mathrm{M}_{t_1}, \mathrm{M}_{t_2})$ and the intrinsic camera parameters (K). Additionally, the stereo correspondences of $\mathbf{p}_{t_{1R}}$ and $\mathbf{p}_{t_{2R}}$ are passed. Rays from all points are calculated by $f_{3\mathrm{D}}(p_t, \mathrm{M}_t, \mathrm{K}) = \mathbf{r}_t^o, \mathbf{r}_t^d$ where $\mathbf{r}_t^o$ and $\mathbf{r}_t^d$ are ray's origin and direction respectively. At each time step, the intersections of stereo-pair rays are computed (green points) and used to guide the ray density during optimisation. Then, samples along the

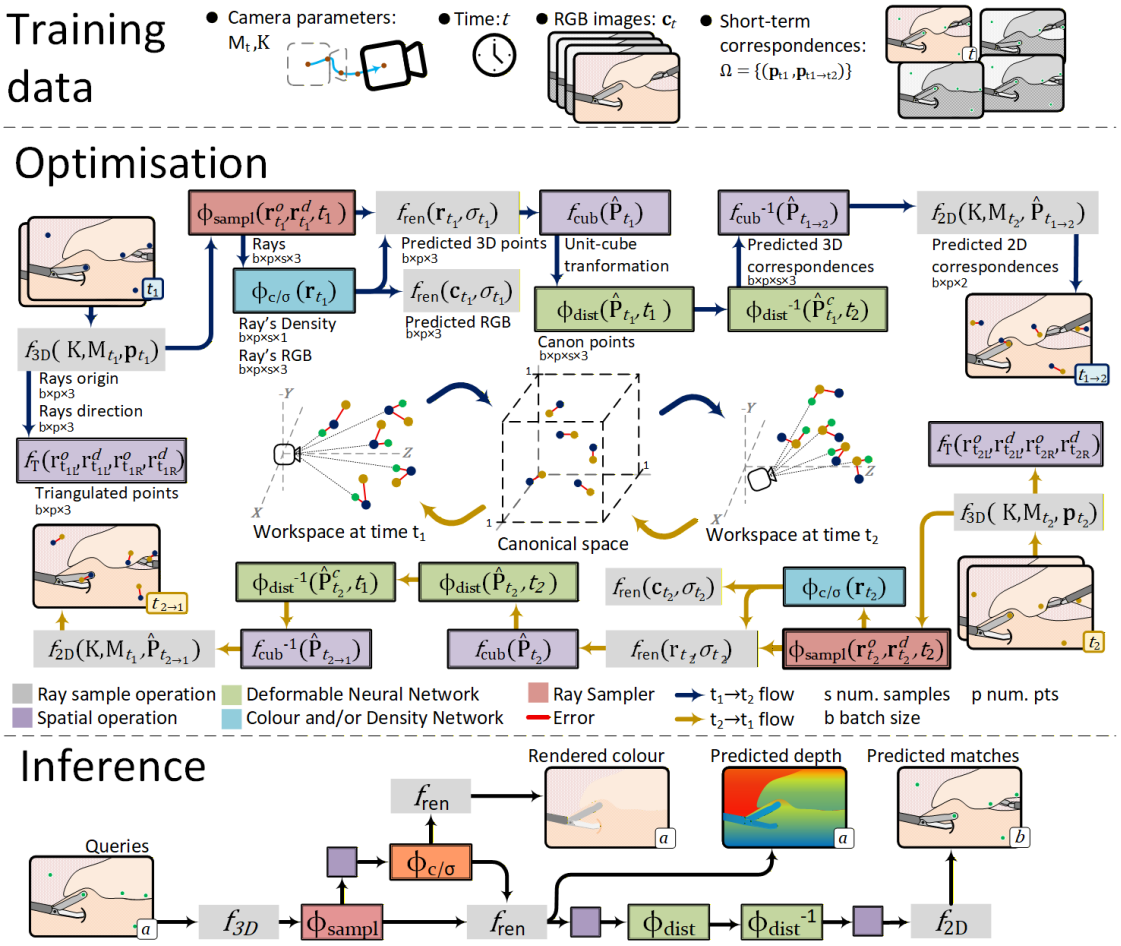


Figure 4.1: **Input.** For each given pair of images from different time steps ($t_1 \neq t_2$), the input is the camera parameters and a set of matched pixel coordinates, with their colours and 2D location on the image planes. **Optimisation.** The information from each time-step flows through the networks bidirectionally ($1 \to 2, 2 \to 1$). Stereo correspondences are used to triangulate 3D points that guide the rays' density (Green points). Rays from the left frames are sampled by ϕ_{sampl}, $\phi_{c/\sigma}$ predicts colour and density for each sample, and the final colour ($\hat{\mathbf{c}}_t$) and 3D location ($\hat{\mathbf{P}}_t$) for each pixel are rendered. These 3D points are mapped from a deformable workspace to a static canonical space ($\hat{\mathbf{P}}^c_t$), then back to the deformable space at a different time step ($\hat{\mathbf{P}}_{t_1\to t_2}, \hat{\mathbf{P}}_{t_2\to t_1}$), and projected to the image plane ($\hat{\mathbf{p}}_{t_1\to t_2}, \hat{\mathbf{p}}_{t_2\to t_1}$). Then the error in each space (red lines), image plane, workspace and canonical space is used to optimise the model. **Inference.** The optimised representation holds all the information of the scene, so queries can be sparse or dense points along with camera pose and time to render an image, depth or find pixel correspondences along the whole video.

rays (only from left frames) are taken by ϕ_{sampl}, each sample is processed by $\phi_{c/\sigma}$ to predict the density σ_t and colour $\mathbf{c}_t$ of each sample. A function f_{ren} calculates the rendering weight from the density values as described in [139] to render the colour

of the pixels $\hat{\mathbf{c}}_t$. Similarly, the locations of the samples are merged into a single 3D point per pixel (blue points $\hat{\mathbf{P}}_{t_1}$ in the workspace at time step t_1 and yellow points $\hat{\mathbf{P}}_{t_2}$ in the workspace at time step t_2). The 3D points then go through a spatial transformation (f_{cub}) that contracts the workspace to a unit cube. The invertible neural network (ϕ_{dist}) performs the bidirectional mapping, considering both spatial and temporal information, between the workspace (deformable) and the canonical space (static). Thus, canonical points ($\mathbf{P}^c_{t_1}, \mathbf{P}^c_{t_2}$) are transformed to the time step opposite to the one from which they came, i.e. blue points in the workspace at t_1 are mapped to the canonical space then to t_2 ($\hat{\mathbf{P}}_{t_1 \to t_2}$), and yellow points in the workspace at t_2 are mapped to the canonical space then to t_1 ($\hat{\mathbf{P}}_{t_2 \to t_1}$). Finally, these points are expanded from the unit cube and are projected to the image planes to obtain the 2D correspondences $\hat{\mathbf{p}}_{t_1 \to t_2}$ and $\hat{\mathbf{p}}_{t_2 \to t_1}$. Data flows in both directions ($t_1 \to t_2$ and $t_2 \to t_1$) during the optimisation to ensure consistency and chaining aggregated correspondences by minimising the correspondence error across all spaces (red lines in image planes, workspaces and canonical space). However, during inference (after optimisation), f_{LT} can be used, providing camera pose and 2D query points at any time step t_1 to render depth and colour, or extract correspondences in the image plane and the 3D space when a target time step t_2 is also provided.

The rest of this section first describes the data processing steps to obtain short-term correspondences, 3D information through triangulation and the tool's mask to simplify the optimisation. Next, we describe the building blocks of the approach, which include a ray sampler, a density and colour predictor and an invertible network. Finally, we discuss our optimisation approach, highlighting two aspects: our multi-scale approach for efficient pixel sampling and our loss function for cross-consistency across the image planes, the deformable space, and the canonical space.

4.2.1 Initial processing

Establishing short-term correspondences

Similarly to [122], pixel correspondences between pairs of images are calculated using RAFT[140], then they are classified as non-reliable, reliable, or occluded in a consistency mask $\text{CM}_{t_1 \to t_2} \epsilon \{0, 1, 2\}$. In this thesis, a small maximum number of pairs per time step and a gap between the pairs of each time step are set to reduce memory usage. To address the limitation of RAFT, we use CoTrackerV3 [141] to recover lost correspondences. We also experiment with correspondences from the Multi-Flow dense Tracker model (MFT)[142], which identifies the most reliable sequences (chains) of flows per pixel to enhance long-term correspondences, and

allows us to use larger time-step gaps between image frames.

Triangulation of target points

Low or unguided ray density in NeRF models is a disadvantage for efficiently exploring rays, identifying regions of space where objects are located, and obtaining pristine density signals (i.e., clear, noise-free). Therefore, our approach leverages the camera's extrinsic parameters and reliable point correspondences between stereo images to guide the density of the rays projected from the camera models. So given pixel correspondences $(\boldsymbol{p}_{tL}, \boldsymbol{p}_{tR})$ between a stereo pair of images, the rays' origin and direction from those pixels are calculated to find the point of intersection. Since the rays from two matched pixels do not necessarily intersect, we define the intersection point as the closest point on the left ray to its matched right ray. In this way, the intersection can be defined by $\mathbf{r}^o_{tL} + \alpha\mathbf{r}^d_{tL} = \mathbf{r}^o_{tR} + \beta\mathbf{r}^d_{tR} + \gamma\hat{\mathbf{d}}_t$. Where $\hat{\mathbf{d}}_t$ is a unit vector along the shortest distance between the rays and the cross product between the ray's direction. This equation can be rewritten in its matrix form and solved as a system of three equations, so the intersection is $\mathbf{P}_t = \mathbf{r}^o_t + \alpha\mathbf{r}^d_t$. During training, the ray density is encouraged to be concentrated around this point. Note that when the unit vector along the shortest distance between the rays is indeterminate, the rays are parallel. We mask out those pixels from the ray density optimisation function, allowing neighbouring pixels to define the density distribution in those regions. This often happens when stereo images are not available, and consecutive frames are used instead. Monocular scenarios are further discussed in the results section.

Workspace construction

As described by Mip-NeRF 360 [143], we define a Axis-Aligned Bounding Box (*aabb*) for our use case that defines the workspace size, we compress the ray's samples into a unitary cube and shift their values to the positive coordinates of the Cartesian axis by using the function $f_{cub}(\mathbf{r}) = (\mathbf{r} - aabb_{min})/(aabb_{max} - aabb_{min})$ which is conveniently invertible (Figure 4.2 a) $\leftrightarrow$ b)). This allows our model to avoid abrupt responses to sign changes. Additionally, we encourage the canonical space to maintain the workspace boundaries by wrapping the unit cube with a unit sphere. In the optimisation section below, we detail how, during training, we bring back all the rays trying to escape from the limits of the unit sphere. This imposes the assumption that all the scenarios processed by our model strictly fit within the boundaries of a defined workspace, which is valid for surgical scenarios, since the abdominal cavity does not grow indefinitely and its dimensions vary only within a limited range across patients, so a single workspace bound accommodates any case.

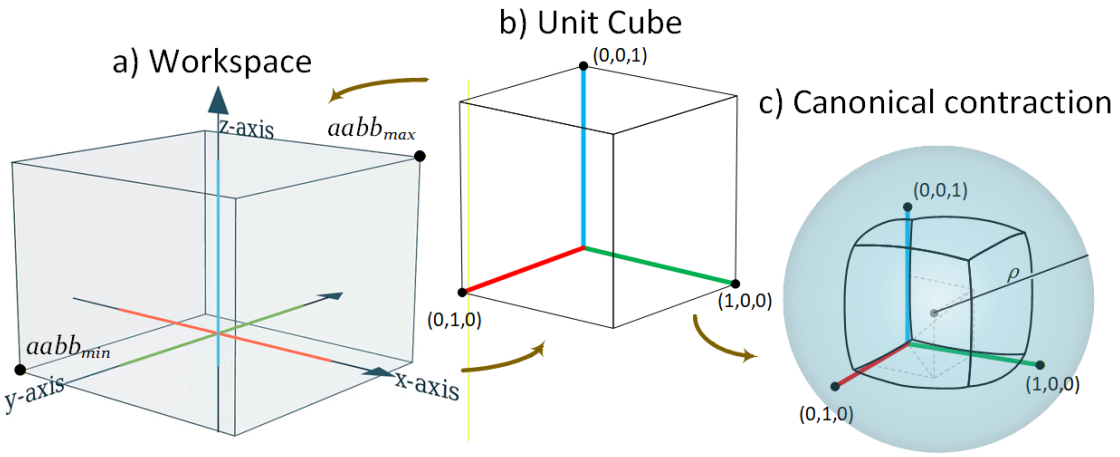


Figure 4.2: **Workspace construction** The transformation of the workspace from a to b is reversible and implemented to improve the efficiency and stability of the NN. The canonical contraction from b to c is done to avoid the divergence of the model due to a bad initialization.

Tool masking

To simplify correspondence aggregation in InvNeRF, we identified the surgical instruments in each image of the video. Using SAM [144] and manual priors on the first image, we segment and propagate labels through the frames of the videos (Figure 4.3). We used this mask to segment the tool during scene optimisation, focusing only on the tissues (deformable objects).

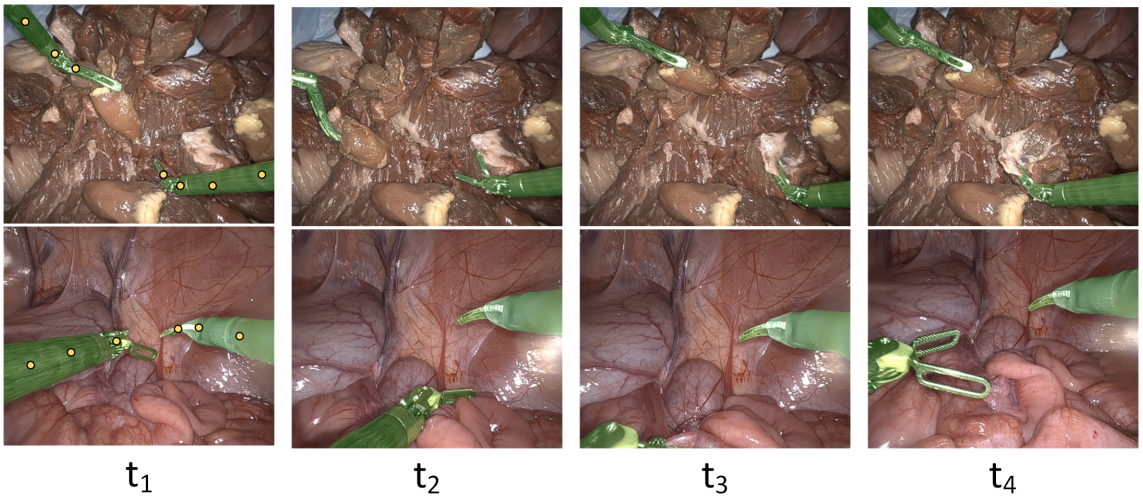


Figure 4.3: **Tool segmentation.** We segmented the tool to focus solely on the deformation of tissues using SAM. The SAM model was fed the videos and manually selected points (the yellow point in the first frame of each sequence) as queries for each video.

4.2.2 Model selection

Ray sampling (Eq. 4.1) is a crucial operation for fast rendering processing. In this work, we compare a fixed and evenly spaced sampling (which is not affected by deformations but requires a large number of samples) against a dynamic proposal network (DPN) to accelerate the process by taking fewer samples. DPN serves as the proposal network from NeRFAcc [139], considering a 4th-dimensional input (after applying the space distortion function f_{cub}, we concatenate the normalised time steps).

$$\phi_{sampl}(\mathbf{r}_t^o, \mathbf{r}_t^d, t) = \mathbf{r}_t \tag{4.1}$$

For the colour and density estimator (Eq. 4.2), we compared the performance of the network used in [122], a GaborNetwork, against our multi-scale HexPlanes (MHP) (Figure 4.4). We built upon HexPlane [145]; we keep the structure, but we concatenate multiple resolutions with a small number of features. In this way, we reduced the number of parameters required to build the model and maintained the ability to capture fine details in density changes. We use two Multi-scale HexPlanes, one to encode the density values and one to encode the RGB values. Low-resolution planes quickly concentrate the density of all the rays to a distance where the object is likely to be located, and high-resolution planes take care of finer details.

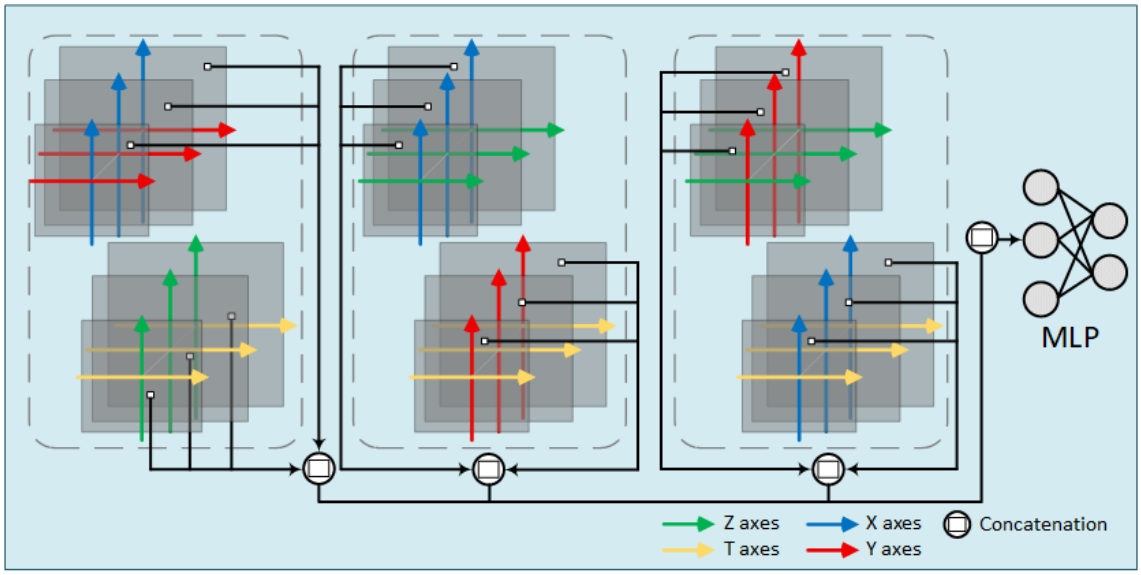


Figure 4.4: **Multi-scale HexPlane**. The 4D coordinates of each pixel (x,y,z,t) are used to sample features from all planes. Features are concatenated and processed by an MLP.

$$\phi_{c/d}(\mathbf{r}_t) = (\hat{\mathbf{c}}_t, \sigma_t) \tag{4.2}$$

The invertible network (Eq 4.3) has three main components: a feature generator (for spatial and temporal features), a network that processes those features and an invertible function that is parametrised by the network's response. We test the feature generators (encoders and feature planes) and invertible functions (CaDeX and CaDeX++) presented in [122] and [138]. However, we removed the use of homography transformations, reduced the number of layers, and used hyperbolic tangent activation functions in the network when using CaDeX.

$$\begin{aligned} \phi_{deform}(\hat{\mathbf{P}}_{t_1}, t_1) &= \hat{\mathbf{P}}^c_{t_1} \\ \phi^{-1}_{deform}(\hat{\mathbf{P}}^c_{t_1}, t_2) &= \hat{\mathbf{P}}_{t_1 \to t_2} \end{aligned} \tag{4.3}$$

4.2.3 Optimisation

This section describes our test-time optimisation process for finding a function f_{LT} parametrised by our invertible NeRF-based architecture, InNeRF, which is run once per sequence at inference rather than as a training stage over a dataset. Along with the architecture that defines f_{LT}, our pixel sampling algorithm and proposed loss function are key components of our framework. The algorithm's objective is to reduce redundant pixel sampling, and our optimisation focuses on global motion estimation, that is, the motion of a point across the whole temporal window rather than between consecutive frames, and geometrically consistent ray density.

Pixel sampling algorithm

Our strategy to avoid redundant pixel sampling is to keep track of the 3D short-term correspondence error for each frame pair in a light error map. Algorithm 2 shows the steps to initialise the error map, to update it, and how to use it when sampling pixels at any iteration. The initialisation starts at the lowest requested scale, images are rescaled accordingly, and the dimensions of the rescaled images are used to initialise the error map (e_{map}) with a fixed cell size for each possible frame pair. The error map is updated every validation iteration, where for every frame pair, one random point within each cell (e_{grid}) is processed by f_{LT}, the error is calculated and used to update the error map. The reduced number of samples allows us to quickly perform this iteration and identify the areas where our function, f_{LT}, needs to focus. The average error over the map ($e_{current}$) is also used to decide when to upsample the size of the error map. The algorithm details how we sample pixels given a selected pair of frames with pixel correspondences based on the calculated error (Figure 4.5). In addition, keeping track of the errors also allows us to consider cases where pairs of frames have already been solved. During optimisation, we select

t based on the error map related to that time step, thereby avoiding redundant frame sampling too. It is important to notice that despite rescaling images and flow maps during the optimisation of the invertible NeRF representation, coordinates in the workspace remain consistent since the camera parameters are modified accordingly to the scales used during the optimisation.

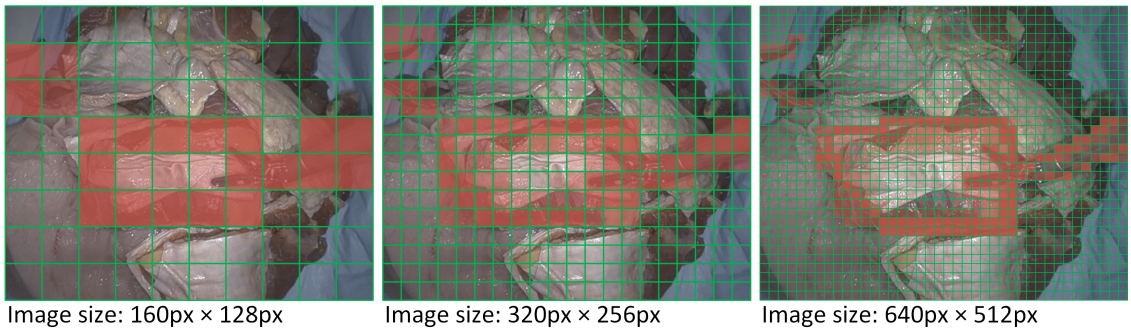


Figure 4.5: **Multi-scale error grid sampling.** Our error grid (green grid) begins at a coarse resolution, with pixel sampling concentrated in cells showing higher local error (red cells). As training progresses, the grid is refined to higher resolutions, focusing sampling on smaller, high-error regions. In the example, the images display the status of the normalised error grid immediately before the resolution is changed.

Loss function

The components of the loss function were categorised by sub-space.

Image plane: We used an RGB loss $\mathcal{L}_{RGB} = ||\mathbf{c}_{t_1} - \hat{\mathbf{c}}_{t_1}||^2 + ||\mathbf{c}_{t_2} - \hat{\mathbf{c}}_{t_2}||^2$; here, the rendered colour from our model is directly compared against the pixel colour on the images. On the image plane, we also evaluate the pixel correspondences between frames with the flow loss [122] $\mathcal{L}_{flow} = ||\boldsymbol{p}_{t_1 \to t_2} - \hat{\boldsymbol{p}}_{t_1 \to t_2}||^2 \mathbb{1}_{\{\mathrm{cm}_{t_1 \to t_2} \in [1,2]\}} + ||\boldsymbol{p}_{t_2 \to t_1} - \hat{\boldsymbol{p}}_{t_2 \to t_1}||^2 \mathbb{1}_{\{\mathrm{cm}_{t_1 \to t_2} \in [1]\}}$. This function masks out non-reliable correspondences using the consistency mask when moving from t_1 to t_2. In the opposite direction, the occluded regions are also masked out because the correspondences are calculated from t_1 to t_2, so the predicted points $\hat{\boldsymbol{p}}_{t_2 \to t_1}$ from the non-occluded pixels in I_2 do not have a correspondence.

$$\mathcal{L}_{2D} = \lambda_{flow} \mathcal{L}_{flow} + \lambda_{RGB} \mathcal{L}_{RGB} \tag{4.4}$$

Dynamic space: Inspired by EndoNeRF [119], and given the limited motion of the camera in a surgical scenario, we incorporate a Gaussian target loss to guide the

Algorithm 2 Pixel sampling algorithm

INITIALISATION
n: number of frames
m: maximum number of pairs per frame
sz: step-size, s: scale
$H \leftarrow s \cdot H, \quad W \leftarrow s \cdot W$
$e_{map} \leftarrow \mathbf{0} \mid \mathbf{0} \in \mathbb{R}^{n,2m,H/sz,W/sz}$
$e_{grid} \leftarrow (i \cdot sz, j \cdot sz) \mid i \in [0, W/sz], j \in [0, H/sz]$
$e_{total} \leftarrow inf$
ERROR MAP UPDATE (DURING VALIDATION)
for $t \in [0, n]$
  for $k \in [-m, m] \backslash \{0\}$ do
    $p_t \leftarrow e_{grid} + \Delta \mid \Delta \sim \mathcal{U}[0, sz]$
    $e \leftarrow \mathcal{L}_{LT}(p_{t \rightarrow t+k}, f_{LT}(p_t, t, t+k))$
    $e_{map}(t, t+k) \leftarrow e$
$e_{current} \leftarrow \text{avg}(e_{map})$
if $e_{current} > e_{total}$ and $s < 1$ then:
  $H \leftarrow 2 \cdot H, \quad W \leftarrow 2 \cdot W$
  $e_{grid} \leftarrow (i \cdot sz, j \cdot sz) \mid i \in [0, W/sz], j \in [0, H/sz]$
  $e_{map} \leftarrow \text{interpolate}(e_{map}, \text{ new shape} = H/sz \times W/sz)$
  $s \leftarrow s \cdot 2$
else:
  $e_{total} \leftarrow e_{current}$
PIXEL SAMPLING (EVERY OPTIMISATION ITERATION)
$P_{t,t+k}(i,j) = \frac{e_{map}(t,t+k,i,j)}{\sum_{i,j} e_{map}(t,t+k,i,j)}$
$\mathbf{p}_t \leftarrow \{(i,j) \mid (i,j) \sim P_{t,t+k}\}$

density of the rays around triangulated 3D target points from stereo images. Given the target points $\mathbf{P}_{t_1}$, rays origin and samples along the rays, we estimate an ideal density around the triangulated depth points, $\sigma_{target} = e^{-(||\mathbf{P}_{t_1} - \mathbf{r}^o_{t_1}|| - ||\mathbf{r}_{t_1} - \mathbf{r}^o_{t_1}||)/2\sigma_{std}}$ and a target rendering weights as $w_{target} = \sigma_{target}/||\sigma_{target}||$. Finally, we minimise the error between the target weights and the rendering weight predicted from the predicted density on the ray's samples using a Kullback-Leibler divergence loss, $\mathcal{L}_{GT} = w_{target}(log(w_{target}) - log(\hat{w}))$.

Similarly to Omnimotion, we also add a smooth flow loss, but in the tracked 3D points, resulting in a smooth displacement along X, Y, and Z, $\mathcal{L}_{smooth} = ||(\hat{\mathbf{P}}_{t_1 \rightarrow t_1+1} - \hat{\mathbf{P}}_{t_1}) - (\hat{\mathbf{P}}_{t_1} - \hat{\mathbf{P}}_{t_1 \rightarrow t_1-1})||$.

$$\mathcal{L}_{3D} = \lambda_{GT}\mathcal{L}_{GT} + \lambda_{smooth}\mathcal{L}_{smooth} \tag{4.5}$$

Canonical space: The sphere loss constrains the canonical object to the size of the deformed workspace. Every canonical sample that is pushed away from this vol-

ume is then dragged back to its non-deformed state, $\mathcal{L}_{CR} = ||\mathbf{P}^c_{t_1} - \mathbf{r}_{t_1}||^2 \mathbb{1}_{||2\mathbf{P}^c_{t_1} - 1|| > 1}$.

As a complementary loss to the geometric consistency we minimise the distance between correspondences in the canonical space since both should represent the same location in this space, $\mathcal{L}_{XC} = ||\mathbf{P}^c_{t_1} - \mathbf{P}^c_{t_2}||^2 \mathbb{1}_{\{\mathrm{cm}_{t_1 \to t_2} \in [1]\}}$.

$$\mathcal{L}_{canon} = \lambda_{CR}\mathcal{L}_{CR} + \lambda_{XC}\mathcal{L}_{XC} \tag{4.6}$$

4.3 Experiments

4.3.1 Datasets

The STIR (Surgical Tattoos in Infrared) dataset [146] is designed for evaluating tissue tracking algorithms using Infrared (IR) markers created with Indocyanine Green (ICG) dye. Videos are captured with a da Vinci Xi surgical robot's stereo laparoscope in both visible and IR modes, showing tissue before and after manipulation. The training dataset includes 572 videos with noisy labels. The validation dataset was released as part of the STIR Challenge 2024 [91]. It includes 60 video clips, recorded at 25 fps, with durations ranging from 0 to 60 seconds, and clean 2D and 3D labels for point tracking.

The SCARED (Stereo Correspondence and Reconstruction of Endoscopic Data) dataset [131] consists of endoscopic videos recorded at 25 FPS from fresh porcine cadaver abdomens using a da Vinci Xi system, paired with depth maps obtained via structured light projection. Although the scenes are static (with no tissue deformation or instrument motion), the camera moves throughout the videos, and ground-truth camera poses are provided for each stereo image pair of stereo images. These data are used to validate the integration of camera pose into the proposed framework and to evaluate 3D depth estimation in this scenario.

The SurgT (Surgical Tracking) dataset [147] was assembled for the MICCAI 2022 challenge of the same name to benchmark soft-tissue trackers, and provides 157 stereo endoscopic videos from 20 clinical cases with calibration parameters, recorded for the most part with a da Vinci stereo laparoscope at 25 Hz and averaging 30 s in duration, with a few videos drawn from earlier collections such as Hamlyn. The videos are split into an unlabelled training subset of 125 videos from 12 cases and a validation subset of 12 videos from 3 cases in which a single keypoint per video is followed through the sequence, 7,835 of its 9,907 frames carrying a stereo label and the remainder holding no box because the point is occluded, out of view or flagged as difficult, and a test subset of 20 videos from 5 cases whose annotations were withheld for the challenge, so that only the 3 validation cases are released with labels. Each

annotation is a keypoint selected on the tissue surface and followed manually through both rectified views of the whole video, from which the bounding box distributed with the dataset is derived by projecting a virtual sphere of 2.5 mm radius around the reconstructed point. The centre of that box is therefore the annotated point itself, and its disparity across the stereo pair places it in three dimensions, so both the pixel and the millimetre errors reported below can be computed on the annotated subset.

The Expanded Comprehensive Robotic Cholecystectomy Dataset [148] provides endoscopic video of ex vivo pseudo-cholecystectomy performed on pig livers with a da Vinci Research Kit by seven surgeons, together with robot kinematics, pedal signals and, in its expanded release, COCO-format annotations of the liver as segmentation masks and of the instruments as keypoints placed on rigid parts of the Large Needle Driver, the Fenestrated Bipolar Forceps and the Permanent Cautery Hook, every frame of certain videos being annotated. The masks delimit the organ in each image without identifying any point on its surface, so the only labels that persist from one frame to the next are the instrument keypoints, and following them measures the localisation of a rigid tool, whose motion is that of the robot arm and is already recorded in the kinematics released alongside the video, rather than correspondence on the deforming tissue. Furthermore, those keypoints are annotated in the image plane alone, with no ground-truth 3D position provided for them, so the millimetre errors reported below could not be computed on them in any case. No annotation in this dataset therefore follows a point on the tissue, and it is excluded from the experiments reported in this chapter.

4.3.2 Evaluation metrics

We report the average end-point error (L^2-norm) between the tracked points and the ground truth across videos. The 2D tracking error is reported in pixels (using the original image resolution), and the 3D tracking error is reported in millimetres (mm). The average training time is also reported. For the comparison vs feed forward methods, we also report the accuracy percentage per threshold (δ_n): δ_{4px}, δ_{8px}, δ_{16px}, δ_{32px} and δ_{64px} on the image plane and δ_{2mm}, δ_{4mm}, δ_{8mm}, δ_{16mm}, δ_{32mm} on the 3D space. Finally, the Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) are reported for the rendered images. These two are measures of rendering fidelity rather than of depth: they compare the rendered image against the reference view and are blind to the scale of the geometry behind it, so a sequence whose depth drifts in scale over time can still score well on both. The geometric accuracy of the model is therefore carried by the 3D end-point errors against the sparse

ground truth, and the image metrics are reported alongside them as an indicator of reconstruction quality only.

The evaluation on SurgT is carried out with the metrics of the challenge rather than those above, since the benchmark scores every method on its own terms. The tracker is initialised at anchor frames spaced roughly fifty frames apart and is run to the end of the video from each of them, and the frames the annotator marked as not visible or as difficult are ignored throughout. Accuracy is the mean intersection over union between the predicted and the annotated boxes, and the 2D and 3D errors are the distances between their centres, in pixels and in millimetres respectively, all three accumulated only while the track is alive: a track is declared lost after ten consecutive frames below an overlap of 0.1 in either view, or beyond 100 mm in three dimensions. Robustness is the share of frames the track survives, counted against every frame in which the target was visible together with every frame in which the tracker answered although it was not. Finally, the expected average overlap averages the overlap of every anchor sub-sequence frame by frame over a fixed window of frames since initialisation, padding with zeros once a track is lost, and it is the score by which the challenge ranks its entries.

4.3.3 Implementation details

The experiments presented in this paper were conducted using NVIDIA L40S 48GB GPUs provided by the HPC service, Aire, at the University of Leeds. We used a batch size of 16 frame pairs, 256 pixels per sample, 32 samples per projected ray, a cell size of 16×16 pixels in the error map, a learning rate of 1×10^{-4}, and the Adam optimiser. An epoch is completed when the model has seen all the images in the video, and we update the error map every 25 epochs. The proposed framework comprises two schedulers and one early-stop criterion that track the average and standard deviation of the error map. The first scheduler upscales the resolution of the error map on a plateau by a factor of 2 (as described in algorithm 2), using a patience of 2 epochs, an improvement tolerance of 0.01, and a starting scale factor of 0.25. The second scheduler was used to update the learning rate when it plateaued, with a patience of 3 epochs, a factor of 0.5, and an improvement tolerance of 0.01. Finally, the stopping criterion is triggered when the scale factor of the error map reaches 1, with a patience of 6 epochs and an improvement tolerance of 0.001.

For SurgT, a model is fitted to each annotated sequence and queried at the pixel of every anchor, returning the corresponding surface point in the remaining frames together with its depth. The centre of the predicted box in the left view is placed on that point, and the centre in the right view follows from the disparity that the

predicted depth implies. No box size is predicted: the width and the height supplied with the anchor are carried forward unchanged for the whole sub-sequence, so that none of the overlap the benchmark measures comes from a box sized to the frame on which it is scored. Frames are taken every second frame of the source video, and the sequences of the first two cases are rectified and staged at half of their resolution before fitting.

4.3.4 Results

A thorough evaluation was conducted to assess various aspects of our approach. First, we present the results of our pixel sampler approach. Then, a comparison of the proposed framework against the state-of-the-art test-time optimisation approaches for 2D tracking is presented. Due to the limitations of Omnimotion [122]'s training speed and memory usage, which takes 22.01 hours to optimise a single sequence against the 2.38 hours of our comparable configuration (Table 4.1), we perform this comparison over a small subset of the original STIR dataset (15 videos with fewer than 175 frames) and a temporal window of 16 consecutive frames as defined by FastOmniTrack [138]. Two further limitations of this approach are examined later in this section: the depth it reconstructs bears no relation to the distance to the actual surface, and its orthographic camera assumption leaves the reconstruction without real-world scale. Here, we also present how the inclusion of the components introduced in the previous section, namely the multi-scale HexPlane representation for colour and density in place of a Gabor MLP, the invertible deformation network in its CaDeX and CaDeX++ forms with either an MLP or a plane-based feature generator, and the depth-proposal pixel sampler, affects our framework's performance. Using the validation dataset from STIR Challenge 2024 [91], we present a comparison, including FastOmniTrack and other feed-forward tracking methods presented in the challenge. Unlike the aforementioned tracking models, our model can render and track the scene at full resolution and receive extrinsic camera parameters from a robotic system; therefore, we further evaluated the accuracy of our model using the scarce 3D ground truth available in the training SCARED dataset. Finally, the framework is evaluated on the SurgT benchmark under the protocol and the metrics of the challenge, and compared against the methods reported in it.

Figure 4.6 shows the optimisation comparison of the model over a video when using the 3D short-term correspondence error map for sampling pixels (Guided) against using the random pixel sampler from other TTO methods (Naive). In the example shown, error-map sampling converges (plateaus) significantly earlier and then stops. In contrast, the other approach continues training longer without reaching the status of the error map in our approach (the red areas in iteration 8000 are visibly more intense than in our approach at iteration 6000), and it stops when the maximum number of iterations is reached. In this way, our approach reduces optimisation time and makes more effective use of the precomputed correspondences. This is particularly useful in surgical scenarios in which the tissue of interest is often the only structure being deformed and appears in a specific portion of the image.

The primary comparison is with other state-of-the-art point-tracking render-

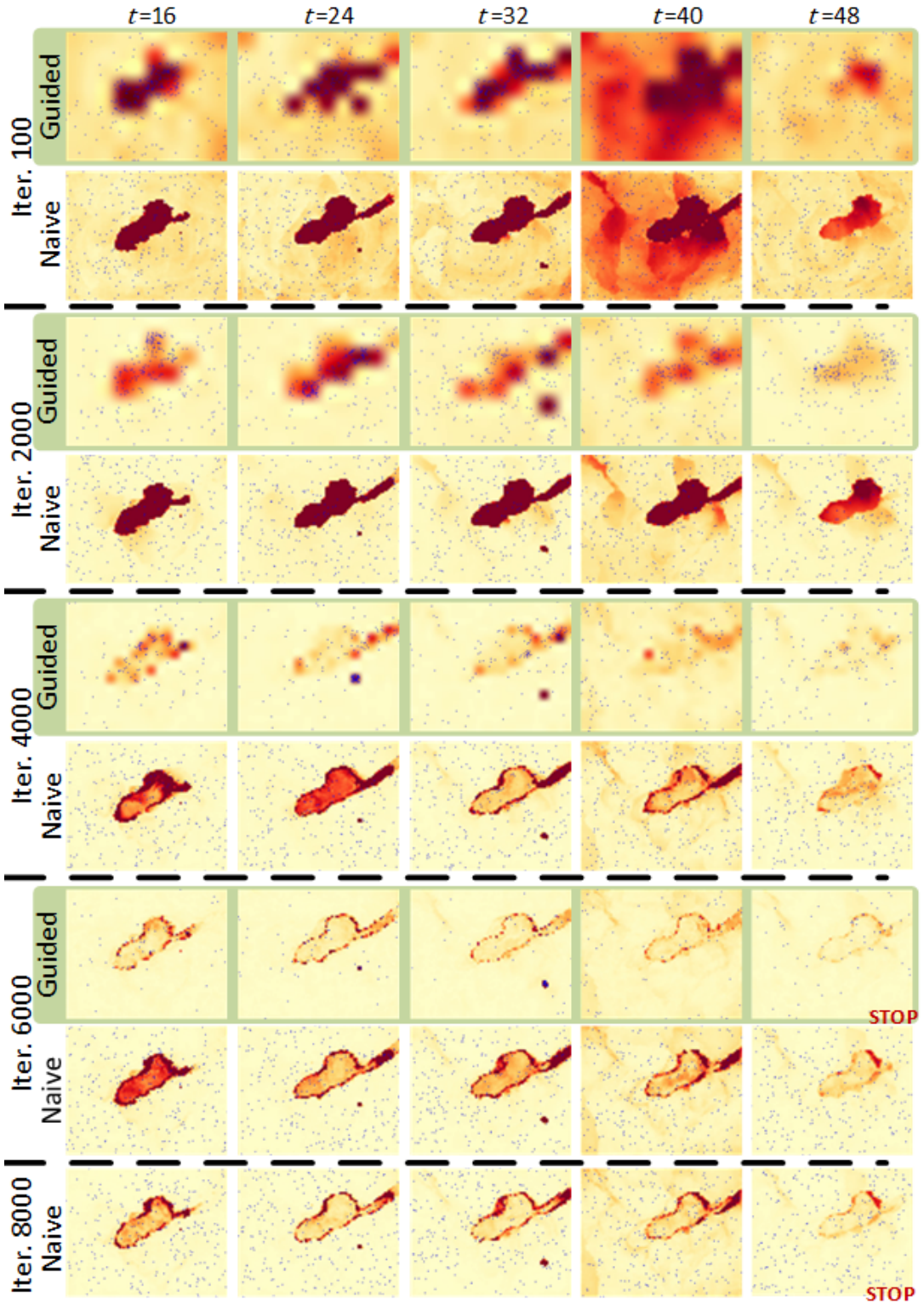


Figure 4.6: **Error map comparison during optimisation** between our guided and a naive approach. Deeper reds signify higher short-term 2D pixel correspondence error. Our guided approach converges with low error at 6000 iterations, while the naive (used by Omnimotion and FastOmniTrack) approach is stopped after 8000 iterations without converging.

ing techniques. Table 4.1 presents the errors of these approaches and our framework across different configurations (discussed in the previous sections). Column ptsB(2D) presents the tracking error over the image plane, whereas column ptsB(3D) and ptsA(3D) present the 3D errors for 3D point tracking and 3D rendered location, respectively. The experiments demonstrate the superiority of our Multi-scale HexPlanes for rendering 3D location in a deformable scenario. They also showed that despite the fact that a proposal network reduces the number of samples in the ray and speeds up the optimisation process, it struggles to accurately concentrate the density of the rays in the areas of interest. Finally, they showed that using a nonlinear function (CaDex++) without additional data structures is the most effective for modelling deformations in a surgical scenario. Overall, the combination of MLP-CaDex++, our MHP, and fixed sampling in the proposed framework (InvNeRF) achieved the best performance. Masking the tools out of the optimisation (InvNeRF$_{(mask)}$) further improves the results of our approach by a 42% in 2D tracking; nevertheless, our approach in all configurations surpasses the TTO SOTA and is the first TTO approach for 3D tracking too. The results in 3D show an average location rendering error of $3.2mm$ and tracking error of $6.2mm$, bringing us closer to achieving consistent and accurate 3D tracking in deformable surgical scenarios. Our approach recovers the benefits of render-based image rendering, and we also report metrics to assess the reconstructed images. We reconstruct the right image from the stereo cameras every 10 time steps and compare it with the original right image (which was not used during optimisation).

Table 4.1: **Test-time optimisation SOTA comparison.**

Model	ϕ_{deform}	$\phi_{c/d}$	ϕ_{sampl}	ptsB(2D) avg l^2-norm	ptsB(3D) avg l^2-norm	ptsA(3D) avg l^2-norm	train-time avg	PSNR	SSIM
Omni [122]	MLP-CaDex	GaborNet	Fixed	31.793_{px}	—	—	22.01_{hrs}	—	—
FOT [138]	Plane-CaDex++	Pred Depth	Fixed	26.863_{px}	—	—	2.00_{hrs}	—	—
Our approach									
InvNeRF	MLP-CaDex	GaborNet	Fixed	17.103_{px}	7.929_{mm}	3.831_{mm}	2.38_{hrs}	21.006	0.509
InvNeRF	MLP-CaDex	HexaPlane	Fixed	17.534_{px}	6.979_{mm}	3.764_{mm}	2.13_{hrs}	**22.496**	**0.556**
InvNeRF	MLP-CaDex	HexaPlane	DPN-32	19.909_{px}	19.639_{mm}	19.109_{mm}	1.80_{hrs}	17.553	0.440
InvNeRF	MLP-CaDex	HexaPlane	DPN-16	22.350_{px}	46.475_{mm}	44.822_{mm}	$\mathbf{1.42}_{hrs}$	17.576	0.441
InvNeRF	MLP-CaDex++	HexaPlane	Fixed	16.919_{px}	6.776_{mm}	3.537_{mm}	2.21_{hrs}	22.489	**0.556**
InvNeRF	Planes-CaDex++	HexaPlane	Fixed	18.255_{px}	7.099_{mm}	3.648_{mm}	1.80_{hrs}	22.479	**0.556**
InvNeRF$_{\text{mask}}$	MLP-CaDex++	HexaPlane	Fixed	$\mathbf{13.925}_{px}$	$\mathbf{6.202}_{mm}$	$\mathbf{3.205}_{mm}$	2.03_{hrs}	22.430	**0.557**

Figure 4.7 shows the difference in the reconstructed depth of the three models; Omnimotion's does not hold any relationship with the depth to the actual surface of the objects (the surfaces blend or are positioned incorrectly with respect to each other), which is probably one of the main reasons it takes so long for the model to find a solution. FastOmniTrack directly utilises precomputed scale-agnostic depth over an orthographic camera model and tracks it, which is more consistent with the object's surface; however, it is unreliable for 3D reconstruction. Our approach has

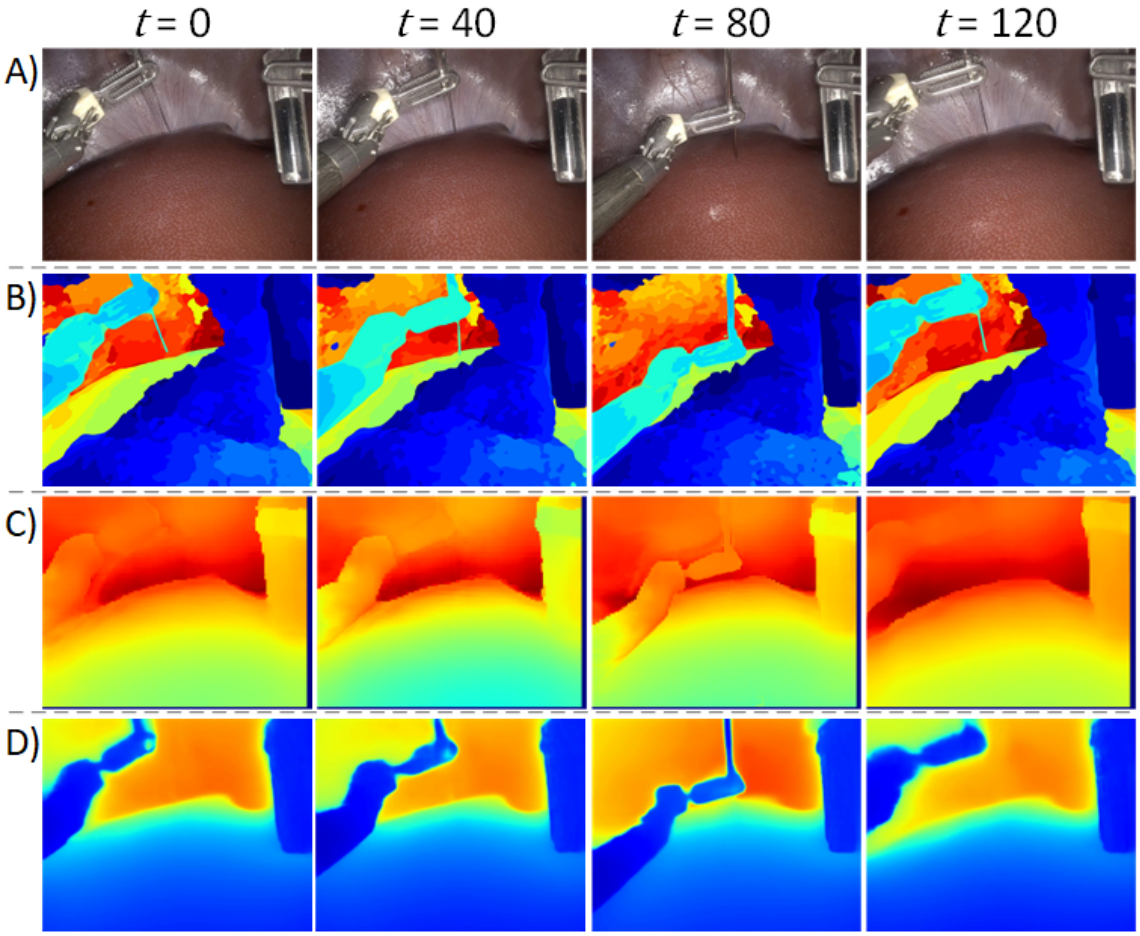


Figure 4.7: **Qualitative depth comparisons of test-time optimisation SOTA.** RGB images (A) and depth maps (B,C,D) at times t =, 0, 40, 80, and 120 s from a sequence. OmniMotion [122] (B) and FastOmniTrack [138] (C) assume an orthographic camera and do not offer real-world scaling. Our approach, InvNeRF (D), uses a pinhole camera and offers a metric-scale depth map.

been demonstrated to be an alternative that utilises a pinhole camera model and a NeRF-based model to aggregate 3D data from a dynamic camera and generate accurate, metric estimates of the object's location.

On the comparison using the full validation set presented in STIR Challenge 2024 [91], we included FastOmniTrack, the feed forward approaches that competed in the challenge, our top performance configuration (InvNeRF with fixed sampler, MHP and MLP-CaDex++), this configuration with masked tools ($\text{InvNeRF}_{\text{mask}}$) and with MFT correspondences ($\text{InvNeRF}_{\text{MFT}}$), Table 4.2. Among the render-based methods, our approach maintains the top performance. With the use of MFT correspondences, we further improve the results, which demonstrate that the TTO approaches will always benefit from every improvement made to obtain reliable correspondences. The accuracy difference relative to MFT may be due to the inclusion

of suboptimal chains on some pixels, which MFT removes during its per-pixel flow analysis. We also observe that, despite FastOmniTrack's strong performance at the lowest threshold, it falls short at the other thresholds, which explains the lower overall average compared with our approach and demonstrates that the proposed approach has greater consistency regardless of the complexity of the deformations the tissues undergo. For 3D tracking (Table 4.3), we include only the different configurations of our framework as rendered-based methods, since FastOmnitrack is not metric. Compared with methods that focus solely on point tracking, our method surpasses them while retaining the advantages of rendering-based methods, including optical flow inference in both 2D and 3D, as well as 3D reconstruction and image synthesis.

Table 4.2: Comparison of InvNeRF against non-rendering based methods on 2D point tracking

Method	δ^4_{px}	δ^8_{px}	δ^{16}_{px}	δ^{32}_{px}	δ^{64}_{px}	δ_{avg}
Baseline						
RAFT	7.26	19.56	39.92	64.92	80.44	42.42
CSRT	22.78	47.38	67.14	74.80	81.05	58.63
MFT	42.54	69.36	86.49	**93.35**	96.37	77.62
Miccai 2024 competitors						
ICVS_2AI	25.40	51.01	74.19	88.71	92.54	66.37
JMEES	26.00	54.44	77.42	91.43	95.36	68.99
MFTIQ[1]	42.34	68.95	85.89	92.14	94.76	76.82
MFTIQ[2]	**44.36**	**69.56**	85.89	91.13	95.16	77.22
CUHK	40.93	68.95	**87.50**	93.15	96.37	77.38
Test-time optimisation approaches						
FastOmniTrack	31.17	51.04	71.54	83.47	90.16	65.48
InvNeRF	27.02	47.81	73.38	86.90	93.65	65.69
$\text{InvNeRF}_{\text{mask}}$	25.91	51.60	77.94	89.99	94.86	68.05
$\text{InvNeRF}_{\text{MFT}}$	34.90	61.96	84.56	91.49	**96.64**	73.39

Figure 4.8 shows some qualitative tracking and depth results on the STIR datasets of challenging scenes where the tissues are occluded (first tracking and depth rows), move fast (second tracking and depth row), so the correspondences drift. We observe that the tracked points are not lost, the drift is small, and the depth information is consistent with tissue motion.

Since the STIR datasets lack camera position information and their experiments had to be run with a static camera, we also tested our model on videos from the SCARED dataset. Figure 4.8 shows the comparison between the ground truth depth and the rendered depth. We evaluate the accuracy of the depth over the pixels with a valid ground truth, having an average 3.05 mm error, and more than 83% of

Table 4.3: Comparison of InvNeRF against non-rendering-based methods on 3D point tracking

Method	δ^2_{mm}	δ^4_{mm}	δ^8_{mm}	δ^{16}_{mm}	δ^{32}_{mm}	δ_{avg}
Baseline						
RAFT-Stereo	13.94	36.16	60.40	79.80	91.51	56.36
Miccai 2024 competitors						
JMEES	25.70	45.40	65.31	81.16	91.01	61.71
ICVS_2AI	27.88	55.15	75.96	91.11	97.58	69.54
Test-time optimisation approaches						
InvNeRF	22.03	49.27	74.01	88.98	**99.16**	66.69
InvNeRF$_{\text{mask}}$	27.19	53.74	**76.87**	91.00	98.92	69.55
InvNeRF$_{\text{MFT}}$	**32.66**	**57.49**	76.51	**92.39**	98.65	**71.54**

the pixels with an error below 5 mm. These results are comparable to those of supervised methods on single frames and videos from the same dataset [131]. The second part of Figure 4.8 shows the qualitative results for tracking and depth over one sequence. We can see that depth is consistent with the portion of ground truth depth provided in the dataset and that this section also matches the tracked tissue on the 2D images.

The framework was also evaluated on the SurgT benchmark. A model was fitted to each sequence of the validation subset, the only part of the dataset whose annotations are released, and queried from the 155 anchors that those sequences define. One qualification has to be made before the numbers are read. The methods reported in the challenge are causal, seeing only the frames that precede the one they predict, whereas the framework presented here fits the whole clip before it is queried, and the expected average overlap together with the two robustness scores are precisely the measures that reward having seen the sequence; the comparison is therefore indicative of the behaviour of the model rather than a placement in the challenge.

On the expected average overlap, the score by which the challenge ranks its entries, the framework reaches 0.371 against 0.314 for the best entry reported on this subset, and it holds the highest robustness of the comparison in both two and three dimensions, 0.770 and 0.873 against 0.724 and 0.758. It does not lead on precision: an accuracy of 0.692 and centre errors of 6.95 px and 6.45 mm sit behind the leading entries, which cluster between 4.5 and 5.6 px. The framework therefore rarely loses the target but does not sit as tightly on it as the methods that lead on accuracy, and the fit over the whole clip accounts for part of the first half of that statement.

The three cases of the validation subset are three different problems, and Ta-

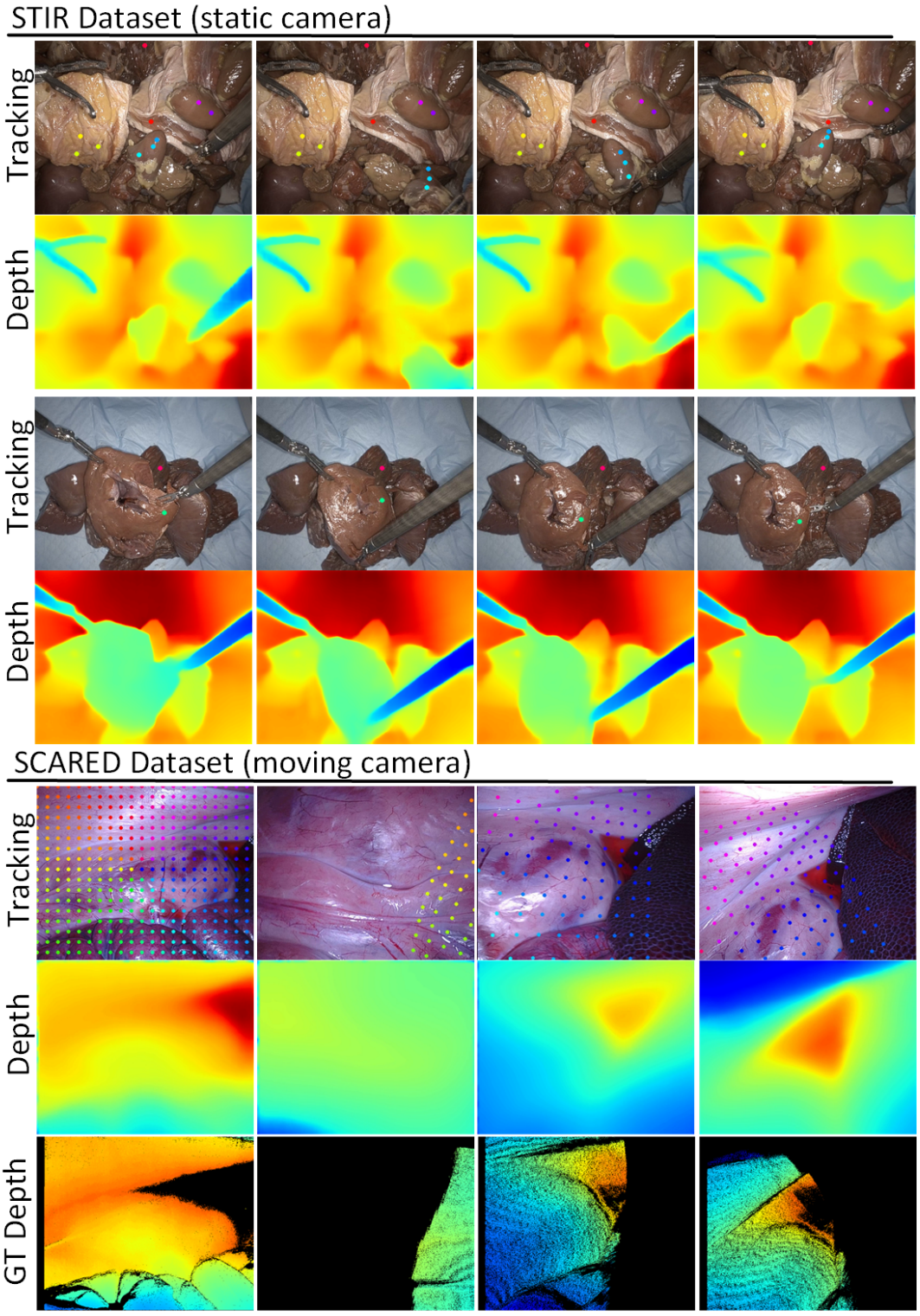


Figure 4.8: **Qualitative results over the STIR and SCARED datasets.** The results on STIR data show the tracking of labelled queried points and the reconstructed depth maps of the shown frames. The results on the SCARED data show the track of a queried grid, the reconstructed depth maps and the ground truth depth maps.

Table 4.4: **Comparison on the SurgT validation subset** against the methods reported in the challenge.

Method	EAO	Accuracy	Rob. 2D	Rob. 3D	Err. 2D (px)	Err. 3D (mm)
InvNeRF	**0.371**	0.692	**0.770**	**0.873**	6.95	6.45
Jmees	0.314	0.795	0.724	0.741	5.5	**2.5**
ICVS-2Ai	0.295	**0.830**	0.684	0.758	**4.5**	6.9
CSRT	0.292	0.766	0.705	0.743	5.6	3.9
ETRI	0.206	0.675	0.635	0.720	14.1	8.6
RIWOlink	0.203	0.698	0.665	0.749	9.5	6.8
SRV	0.131	0.686	0.397	0.721	11.7	11.9
MEDCVR	0.110	0.614	0.436	0.638	7.0	18.4
KIT	0.103	0.672	0.375	0.698	12.8	17.2
TransT	0.095	0.528	0.410	0.673	13.1	23.1

Table 4.5: **Results per case on the SurgT validation subset.** The EAO is defined over the merged curve of the whole subset and is therefore not separable by case.

Case	Accuracy	Rob. 2D	Rob. 3D	Err. 2D (px)	Err. 3D (mm)
Case 1	0.681	0.999	0.999	6.80	1.22
Case 2	0.591	0.455	0.731	12.89	20.57
Case 3	0.800	0.691	0.793	2.81	3.95
Validation subset	0.692	0.770	0.873	6.95	6.45

ble 4.5 separates them. The first is a near-static scene observed by a moving camera, and it is close to solved: the track is never lost over its 61 anchors and the 3D error is 1.22 mm. The second is deforming tissue observed by a moving camera, with the highest velocity in three dimensions of the subset at 3.22 mm per frame and 28% of its frames flagged as not visible or difficult. In this scenario,the framework breaks down, holding the target in fewer than half of the frames for which it is answerable and reaching 20.57 mm. The third is recorded at 480×270, is affected by blur and smoke, and carries the largest proportion of unusable frames of the three at 33%, so its 2D error of 2.81 px, the lowest of the three, says less than it appears to: a pixel there covers more than twice the tissue in each direction that a pixel of the first two cases does, and the hardest sequence of the case follows a target of some 15×15 px at a working distance of 152 mm, where a small error in disparity becomes a large error in depth. Furthermore, and unlike the STIR sequences, which carry no camera pose and were therefore run with a static camera, two of these three cases are recorded by a camera that moves throughout, so the evaluation now extends to camera motion over real surgical video rather than over the static ex vivo scenes of SCARED. Figure 4.9 shows qualitative results for one sequence of each case.

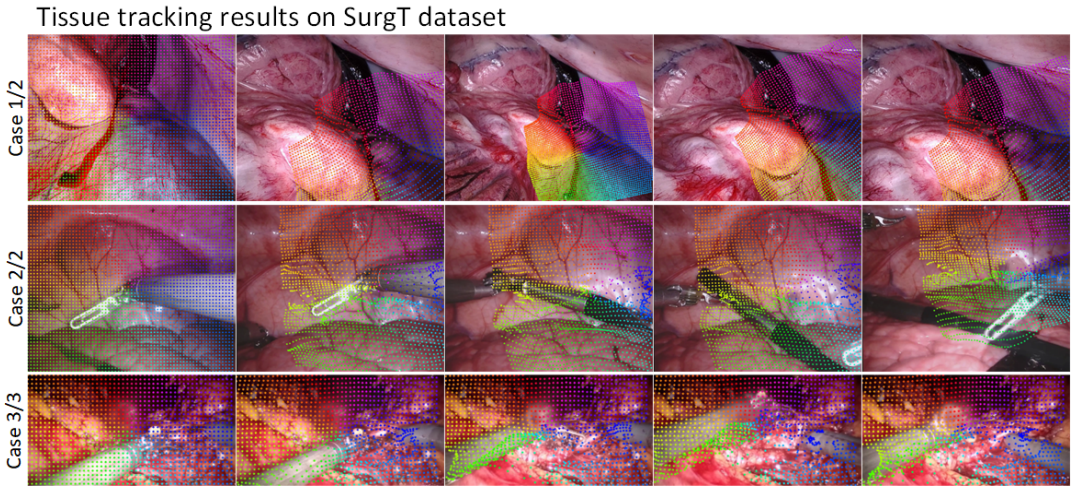


Figure 4.9: **Qualitative results over the SurgT dataset.** One sequence of each case of the validation subset, with frames sampled across the sequence. The overlaid points are queried on the tissue in the first frame and propagated by the fitted model.

4.3.5 Ablation

Table 4.6: Ablation study

GTL	CCL	SFL	PSNR	ptsB(2D) avg l^2-norm	ptsB(3D) avg l^2-norm	ptsA(3D) avg l^2-norm	Time avg
	✓	✓	20.319	32.201	47.538	50.441	**0.855**
✓			**22.453**	18.170	9.736	3.252	1.957
✓	✓		22.446	14.421	6.582	**3.168**	1.958
✓		✓	22.430	18.011	8.312	3.222	1.815
✓	✓	✓	22.429	**13.925**	**6.202**	3.206	2.035

Architecture design. This section provides additional details on the architectural design choices explored before the final model presented in Chapter 4. The goal of these experiments was to analyse how the ordering of deformation, radiance estimation, and rendering operations affects both reconstruction quality and tracking performance. Figure 4.10 illustrates the three architectures evaluated. Each configuration was designed to isolate specific modelling assumptions about where to encode geometry, appearance, and temporal variation.

Model A: Canonical Radiance Field (OmniMotion-style Baseline). The first configuration closely follows the architecture proposed in OmniMotion[122], with the only modification being the camera projection model used for ray generation. In this formulation, sampled points are first mapped to a canonical (static) space, where a colour–density network (Eq 4.2, $\phi_{c/d}$) predicts radiance and volume density. While this approach achieved stable optimisation and reasonable recon-

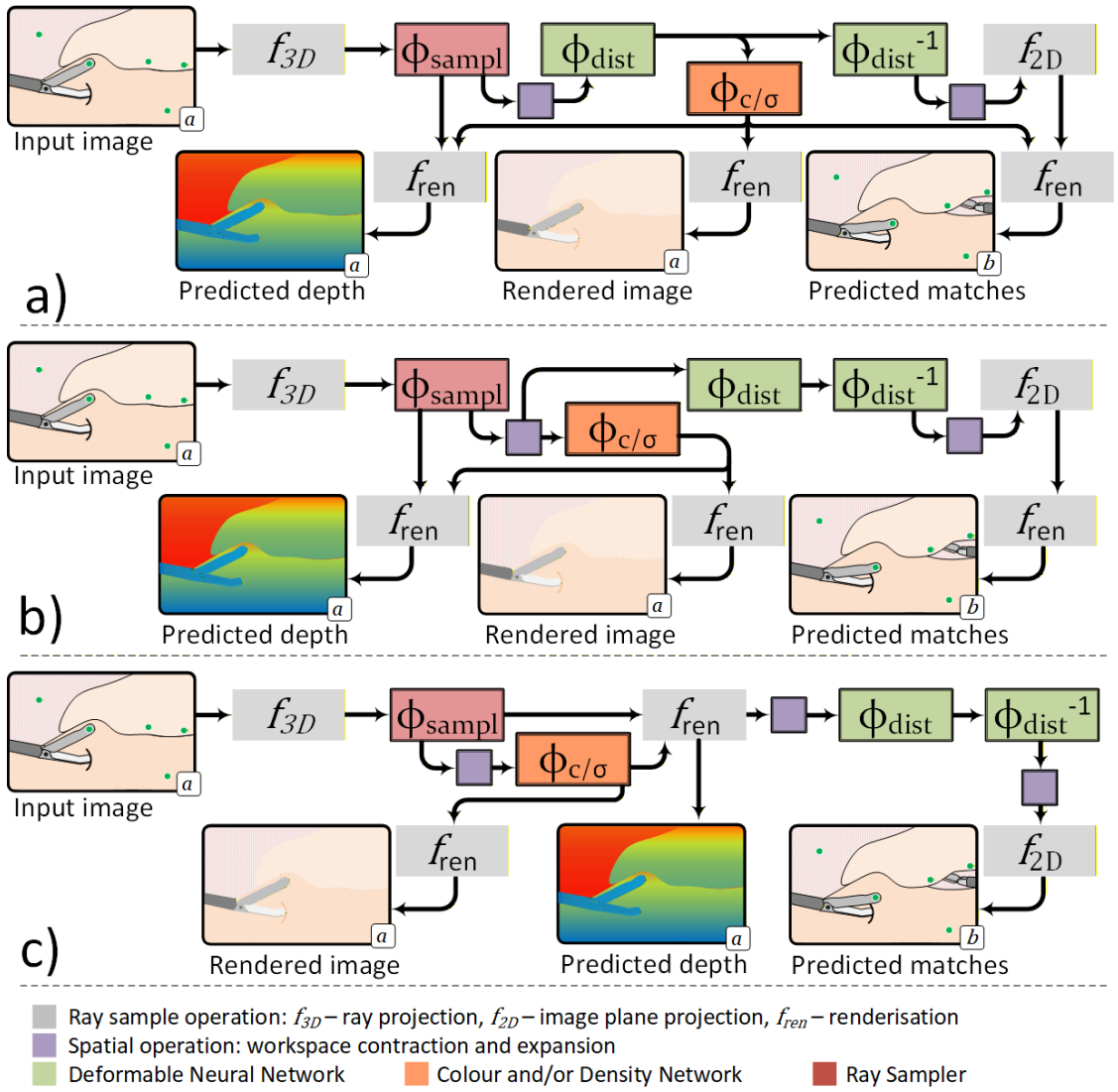


Figure 4.10: **Proposed Invertible Architectures**.

struction quality, it imposes a significant representational burden: a single point in canonical space must encode not only intrinsic tissue appearance and density, but also view-dependent reflectance effects and illumination changes across time. In other words, colour and density predictions are implicitly required to absorb temporal and photometric variability. This coupling of geometry and time-dependent appearance limits disentanglement and reduces modelling flexibility, particularly in highly specular surgical scenes where illumination and viewing angle vary substantially.

Model B: Radiance Estimation in Deformable Space. To alleviate this limitation, the second architecture computes colour and density directly in the deformable space at the sampled point's initial position. In this configuration, time is explicitly provided to the radiance network, enabling the model to condition its pre-

dictions on temporal state. This modification introduces two important advantages: temporal conditioning of appearance and reduced canonical overloading. Which means that $\phi_{c/d}$ explicitly accounts for view-dependent reflectance and illumination changes, and the canonical space no longer needs to encode appearance variation induced by camera motion.

Model C: Rendering Block Before Deformation (Final Model) Similar to the fast-tracking framework [138] but with our camera formulation, the third configuration introduces a rendering block prior to the deformation stage. In this design, $\phi_{c/d}$ is responsible exclusively for identifying and modelling the tissue surface while the invertible deformation network (Eq. 4.3,ϕ_{deform}) encodes temporal geometric transformations. Hence, rendering supervision is partially decoupled from deformation estimation.

All configurations were trained and evaluated on a subset of the STIR dataset under identical experimental conditions. Table 4.7 summarises the quantitative comparison. The results demonstrate a consistent improvement with each architectural modification, supporting the progressive design decisions described above. Importantly, the architectural evolution was not heuristic but guided by empirical evidence and by theoretical considerations regarding the disentanglement of appearance and motion.

Table 4.7: **Quantitative results for tested InvNeRF architectures.**

Model	ptsB L2	pts3D B	pts3D A	mask IoU	time
A	11.140	0.205	0.144	**0.207**	2.59 hrs
B	12.021	**0.153**	0.177	0.162	2.64 hrs
C	**8.507**	0.175	**0.087**	0.197	**2.42 hrs**

Concurrently with this work, Endo-TTAP [149] was proposed for tissue point tracking and also evaluated on the STIR Challenge 2024 data, reporting an averaged 2D accuracy of 78.4 under the same protocol used here, ahead of the best method in Table 4.2. It is not included in the comparison for two reasons. It reports only the averaged 2D accuracy and the endpoint error on this benchmark, without the per-threshold breakdown of Table 4.2 and without any 3D metric, which are reported for SurgT instead and therefore cannot enter Table 4.3. Furthermore, it is a supervised method fine-tuned on the 576 training videos of the same challenge, whereas the framework presented here is optimised at test time without training on the benchmark. A comparison under a common protocol is the natural next step rather than one that the present tables can carry.

Finally, the ablation study over our loss function is presented in Table 4.6, which evaluates the influence of the different components in our loss function: the Gaussian

target loss (GTL), the cross consistency loss (CCL) and the smooth flow loss (SFL). The inclusion of all the components in our loss demonstrated the best trade-off to obtain reliable 2D and 3D tracking within a reasonable time.

Colonoscopic datasets were not included in the main body of the thesis due to the lack of GT correspondences in public datasets, and because colonoscopy falls outside the abdominal laparoscopic scope defined in Chapter 1. However, Appendix B.1 presents complementary experiments on coloscopic scenarios to further evaluate the proposed method, test its use when only a monocamera is available and explore the extraction of an explicit 3D reconstruction from the proposed InvNeRF architecture.

4.4 Discussion and conclusion

We present the first NeRF-based test-time optimisation to address 2D and 3D tracking simultaneously and the first algorithm to avoid redundant pixel sampling on TTO, making the optimisation of surgical scenes more efficient. In the surgical context, our method outperforms any other 2D tracking TTO method, reducing the mean average 2D error by 50% (Table 4.1), and it does not fall far behind other feed-forward 2D tracking approaches (Table 4.2). It also outperforms all methods on 3D tracking by 5% on the average end-point-error (Table 4.3) and retains the advantages of rendering-based techniques. Thus, optical flow in both 2D and 3D, as well as 3D reconstruction of dynamic and deformable scenarios, is encoded in the implicit representation of our InvNeRF (Figure 4.7). A key distinction of our method from other NeRF-based approaches is that we only need to render each new visualised location once, after which we track it. This advantage removes many of the artefacts produced in other approaches.

Our framework maintains a consistent relation with the deformable space by constraining the workspace size and linking 3D correspondences to a small volume in the canonical space. However, the topology of the objects and the geodesic distances in the canonical space could be improved in future works. This essentially could improve the convergence and ability to track points under extreme complex deformations or prolonged periods of occlusion. Finally, as for other TTO methods, any improvement on 2D correspondences will also further improve the representations created by our InvNeRF method.

In conclusion, our approach establishes the state of the art with large improvement in 3D point-tracking compared to existing techniques in surgical scenarios. This is the first method that encodes representations from structure, colour and 3D flow of the object making it applicable towards estimation of accurate points track-

ing and related depth estimation. We demonstrate that our results are comparable to those of methods that specialise in only one of these tasks.

Chapter 5

Spatially and Metrically Grounded Representations

Within the visual perceptual framework introduced in Chapter 2, spatial and metric grounding represents a critical transition from relative, image-based descriptions of the surgical scene to geometrically meaningful representations expressed in a physical coordinate frame. Specifically, estimating depth, scale, and camera pose allows the scene elements to be located and related in three-dimensional space. While relative depth can be recovered from local image primitives within a single frame, metric depth requires resolving scale through geometric constraints arising from stereo vision, camera motion, or known system geometry. In robotic minimally invasive procedures, camera pose information may be partially or fully available through kinematic sensing, providing an opportunity to anchor visual observations in a coherent spatial frame. In this thesis, the focus is on enabling and evaluating metrically accurate depth estimation under such conditions, laying the foundation for spatial reasoning and downstream integration.

Despite extensive research on depth estimation and three-dimensional reconstruction, achieving reliable metric spatial grounding in surgical environments remains challenging. Soft tissue undergoes complex, non-rigid deformation driven by physiological motion and surgical manipulation, while endoscopic imagery is affected by specular reflections, variable illumination, occlusion, and weak texture. These factors violate the assumptions underlying many classical geometric methods and

This chapter is based on the following work: Loza-Galindo, G., Magro, M., Calmé, B., Hu, J., Ruffaldi, E., Jones, D., De Momi, E., Ali, S. and Valdastri, P., *A Dataset for Depth in Robotic Endoscopy with Dynamic Scenarios (DRENDS)*. 2026. Scientific Data. https://www.nature.com/articles/s41597-026-07775-2. DOI: 10.5281/zenodo.17598453. Project page: STORMLAB-DRENDS.

complicate learning-based approaches. In practice, most existing datasets simplify the problem by assuming rigid scenes or static tissue, limiting the evaluation of depth estimation methods under realistic deformable conditions. As a result, the scarcity of dense, scale-accurate depth ground truth in dynamic surgical scenarios continues to constrain progress in metric spatial reasoning and makes it difficult to assess how well existing methods generalise to real procedures.

This chapter addresses this gap by introducing a novel dataset that provides high-resolution endoscopic image sequences with dense, scale-accurate depth ground truth under tissue deformation induced by instrument manipulation. By capturing both phantom and ex vivo organs deformed by instrument probing and retraction, the dataset enables systematic evaluation of metric depth estimation methods in conditions that better reflect intraoperative reality. In addition, this chapter presents a benchmark study over the proposed dataset, highlighting the limitations of state-of-the-art learning-based approaches when applied to dynamic deformable scenes. Rather than proposing a new depth estimation method, this chapter's contributions enable the development, validation, and comparison of metrically grounded scene representations that are coherent with the physical surgical environment. Such representations are essential for geometric reasoning, integration with robotic kinematics, and the construction of spatially consistent models of the surgical scene in subsequent perceptual stages.

5.1 Problem definition

We address the lack of suitable data resources that enable the development, evaluation, and comparison of depth estimation methods under realistic intraoperative conditions.

Formally, given a sequence of endoscopic images acquired during a surgical procedure, the objective in this Chapter is to recover depth representations that are (i) scale-accurate, and (ii) temporally consistent under non-rigid tissue deformation. The approach presented in this section, the Depth in Robotic Endoscopy with Dynamic Scenarios (DRENDS) dataset, is designed to enable metric spatial grounding in deformable surgical scenarios. By pairing high-resolution stereo endoscopic image sequences with synchronised, scale-accurate depth measurements obtained from a calibrated time-of-flight sensor, DRENDS dataset provides per-frame metric depth under controlled yet realistic tissue motion. The dataset further includes precise temporal synchronisation, camera calibration parameters, and an open processing pipeline to ensure reproducibility and facilitate method development.

5.2 Methods

The DRENDS dataset comprises synchronised stereo videos and corresponding ground-truth (GT) depth information that capture tissue deformations during manipulation. To our knowledge, this is the first dataset that includes high-resolution videos of dynamic surgical scenes (i.e., tissue deformed by instrument manipulation) with ground truth depths. The methodology for generating the dataset is organised into four parts: setup design, calibration, acquisition protocol, and data processing.

5.2.1 Setup

The dataset was acquired at the University of Leeds using the da Vinci Research Kit (dVRK) with the da Vinci model Si, Intuitive Surgical, Sunnyvale, CA, USA. The hardware setup consisted of: (i) one da Vinci robotic arm, (ii) an Omega7 haptic device (from Force Dimension, inc. Nyon, Switzerland) as tool manipulator, (iii) the stereo endoscopic camera from the da Vinci system (Ikegami 720p cameras in the phantom setup and 1080i Panasonic cameras in the ex vivo setup), and (iv) a Helios2$^+$ ToF camera from LUCID Vision Labs Inc. Burnaby, Canada. Figure 5.1 illustrates the configurations for the phantom and ex vivo data collections. The stereo and ToF cameras were mounted to maximise the coverage workspace, with the stereo camera positioned so that it did not obstruct the Helios2$^+$ camera view. All devices were interfaced through the Robot Operating System (ROS). During data acquisition, an operator manipulated the tissues with the robotic tool via the tool manipulator device.

5.2.2 Calibration

The intrinsic parameters of the Helios2$^+$ ToF camera were obtained using the Arena Software Development Kit provided by LUCID Vision Labs. The intrinsic parameters of the stereo endoscopic cameras were estimated with images of a planar chessboard calibration pattern (Grid size: 3×4 squares, Square size: 15×15 mm.), using the OpenCV library with the Python interface. Leveraging Zhang [150]'s method for camera calibration first, individual monocular calibrations for the left and right cameras were performed, optimising the intrinsic matrices and distortion coefficients using a nonlinear least-squares algorithm with a termination criterion of 500 iterations or $\epsilon = 10^{-8}$. Initial guesses for the focal length and principal point were set from the image dimensions, and intrinsic parameters were refined while distortion terms remained unconstrained. In the second stage, stereo calibration was carried

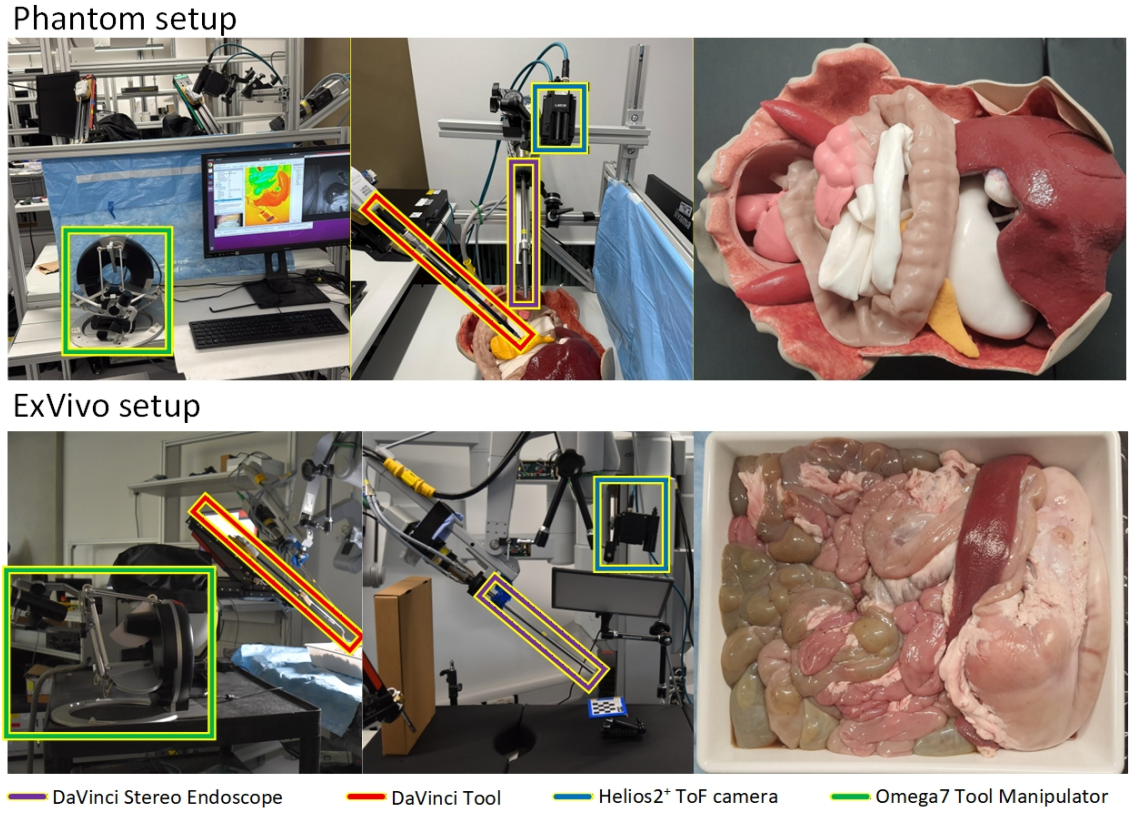


Figure 5.1: Phantom and ex vivo setup, first and second row, respectively. The operator controls the DaVinci tool using the Omega7 device as a tool manipulator; the endoscopic and helios camera views are displayed on a screen next to him and are used for visual feedback during manipulation.

out using the previously estimated intrinsics as fixed initial values and refining the rotation ($\mathbf{R}_{stereo}$) and translation ($\mathbf{T}_{stereo}$) between the cameras. Stereo rectification was applied to all images using the computed projection matrices.

To align the two sensing modalities, the calibration pattern was also captured with the ToF camera, and we estimated the rigid transformation between their coordinate systems using a nonlinear least-squares optimisation in Python. Given 3D points in the depth-camera frame, $\mathbf{X}_i^{\mathrm{PC}} \in \mathbb{R}^3$, and corresponding RGB image points $\mathbf{u}_i = \left[u_i, v_i\right]$, we estimate the rotation $\mathbf{R} \in SO(3)$ and translation $\mathbf{T} \in \mathbb{R}^3$ that map depth-camera points into the RGB camera frame. The transformation of the points is given by $\hat{\mathbf{X}}_i^{\mathrm{RGB}} = \mathbf{R}\mathbf{X}_i^{\mathrm{PC}} + \mathbf{t}$, and the projection onto the image plane is $\hat{\mathbf{u}}_i = \pi(\mathbf{K}, \mathbf{R}\mathbf{X}_i^{\mathrm{PC}} + \mathbf{t})$ where $\pi(\cdot)$ denotes the pinhole projection. Finally, the calibration solves the optimisation described by the equation 5.1. Calibration accuracy was assessed through both the mean 2D reprojection error (in pixels) and the 3D alignment error (in millimetres), with detailed results reported in the Technical Validation section.

$$(\mathbf{R}^\star, \mathbf{t}^\star) = \arg\min_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{N} \left\| \mathbf{u}_i - \pi\big(\mathbf{K},\, \mathbf{R}\mathbf{X}_i^{\mathrm{PC}} + \mathbf{t}\big) \right\|_2^2 . \tag{5.1}$$

5.2.3 Acquisition protocol

Each organ — colon, intestine, stomach, liver, and pancreas — was recorded under three illumination conditions (high setting the da Vinci light source to the 80% of its power, standard setting the da Vinci light source to the 60% of its power, and external, using a white light source parallel to the camera view of the endoscope). A high-illumination camera created specular highlights that reduced visible texture; standard illumination represented optimal viewing conditions; and external illumination was used to recreate scenarios where illumination is limited, yet the scene remains visible. Two acquisition strategies were used: sequence and video. In the sequence approach, 50 non-consecutive frames are captured, each corresponding to a distinct tissue deformation, allowing a broader range of deformation states to be recorded. In the video approach, frames are recorded continuously at 16 FPS while tissues are manipulated through a continuous range of deformations. The two acquisition strategies are performed independently, and the sequences are not subsampled from the videos.

Phantom setup: Data were collected using the synthetic abdominal organs of the EasySurg Training System from XCorpa medical simulators. Together, 15 sequences and 15 videos of 90 seconds, for a total of 23,250 frames. The stereo and depth cameras were synchronised using the timestamps of the computer subscribed to all the ROS nodes and the ROS-Python time synchroniser function.

Ex vivo setup: Data were collected using porcine ex vivo tissues (without liver). Together, 12 sequences and 12 60-second videos add up to 12,120 frames. To handle the interlacing of the 1080i camera, a simple deinterlacing function was built for the endoscopic image publisher, dropping one channel and interpolating from the vertical neighbours' pixels. The acquisition software was built in ROS. This setup relies on the presynchronisation of the clocks in the capture computers, so a post-collection routine was run to filter videos at 16 FPS and ensure synchronised frames from all cameras.

The following protocol was followed in the collection of each sequence:

1. Position the synthetic organ within the phantom framework.

2. Adjust the phantom so that the target organ is centred in the endoscopic view.

3. Bring the surgical tool into the field of view.
4. Set the illumination condition.
5. Test the robotic system connectivity. The trained operator verifies that the tool moves correctly within the camera's view.
6. Start recording.
7. Tissue manipulation:
   (a) During video collection, the operator explores the tissue surface, going from one tissue deformation to another and maximising the visible surface.
   (b) During sequence collection, the operator positions the tissue under different deformations and stays still until a second operator manually triggers the capture of all cameras.
8. Stop recording.
9. If there are consecutive frames in a video with a time-step difference greater than 0.1875 seconds during a tissue manipulation, the video was discarded and re-taken.

5.2.4 Data Processing

To generate the ground-truth depth maps, we applied a multi-stage processing pipeline to the raw point clouds from the ToF camera. The main steps included noise filtering to handle inconsistencies and invalid pixels, masking occlusions, and refining coordinate transformation to align the point cloud with the stereo camera view.

Noise filtering

We start by removing the null and infinite values from the raw ToF point clouds, which were replaced by interpolation from the nearest neighbours. Then a temporal smoothing was applied to depth values, using a sliding window of ± 1 sec. ($\approx$16 frames). Pixel values were smoothed only if depth differences fell within the tolerance error reported by the Helios2$^+$ manufacturer (± 4 mm), ensuring that true depth changes due to tissue motion were preserved. Finally, a spatial filter was used; each frame was smoothed using a bilateral filter[151] (OpenCV implementation) on

a sliding window of 12×12 pixels, which preserves depth discontinuities while reducing local noise. The parameters used for temporal and spatial filtering were empirically selected based on qualitative evaluation of representative sequences, following common practice in depth-map denoising to balance noise reduction with preservation of depth discontinuities and deformation-induced geometric changes. Filtered point clouds were used to generate surface meshes of the scene. Figure 5.2 shows an example of the mesh with and without the filters applied to the point cloud.

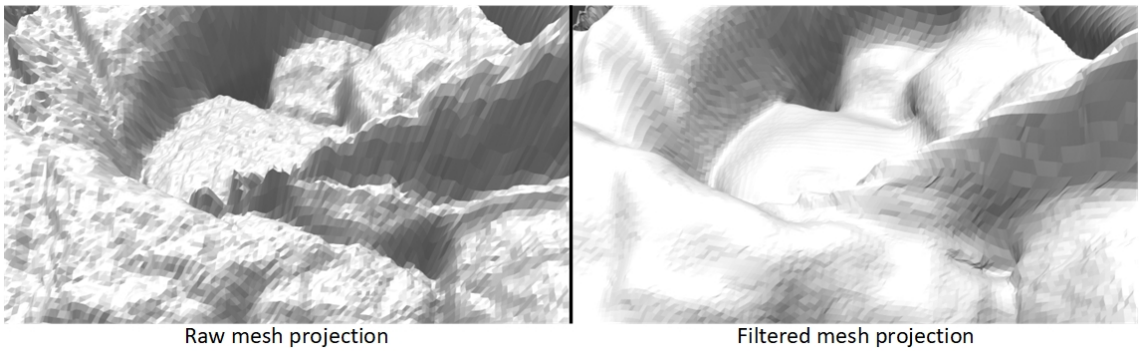


Figure 5.2: Comparison of the reconstructed mesh before and after the noise filter.

Occlusion mask generation

Invalid pixels due to occlusion were identified based on two criteria. First, surgical tools were masked out because their reflective, thin surfaces led to unreliable depth values. Tools were segmented in each camera view using the Segment Anything Model v2[152] (SAM2), with manually selected image points as prompts (Figure 5.3 first and second columns). The prompts used in our pipeline are included in the resources shared with our dataset to ensure reproducibility. Second, to account for viewpoint differences between the Helios2$^+$ and the endoscopic cameras, we removed surface elements that are not visible from the endoscopic perspective. To achieve this, surface normals were computed for each vertex of the Helios mesh before transformation. Vertices whose normals were perpendicular to the Helios z-axis were identified as belonging to surfaces that fall outside the endoscopic field of view and were therefore considered occluded. These regions were subsequently removed from the final depth maps (Figure 5.3, third column).

Coordinate transformation

The transformation from equation 5.1 on the meshes from the ToF camera to the stereo camera-coordinated systems demonstrates some misalignment when depth is

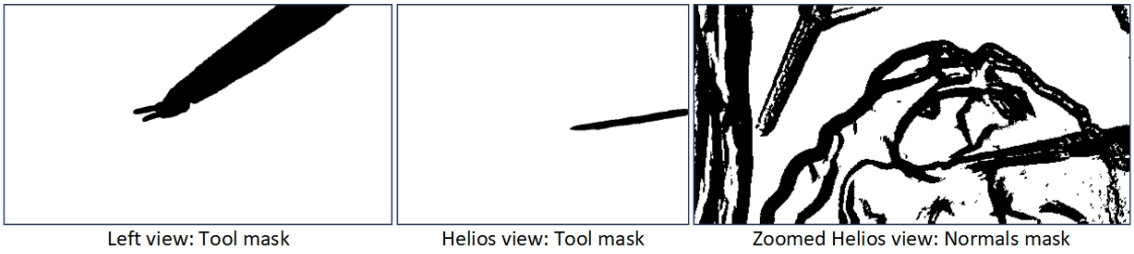


Figure 5.3: Mask to filter out valid surfaces on the meshes that generate depth ground truth. Tool masks in the first and second columns. Normals mask in the third column, it leaves out points that do not capture the location of the surface on the laparoscopic camera view because they were occluded on the Helios camera view.

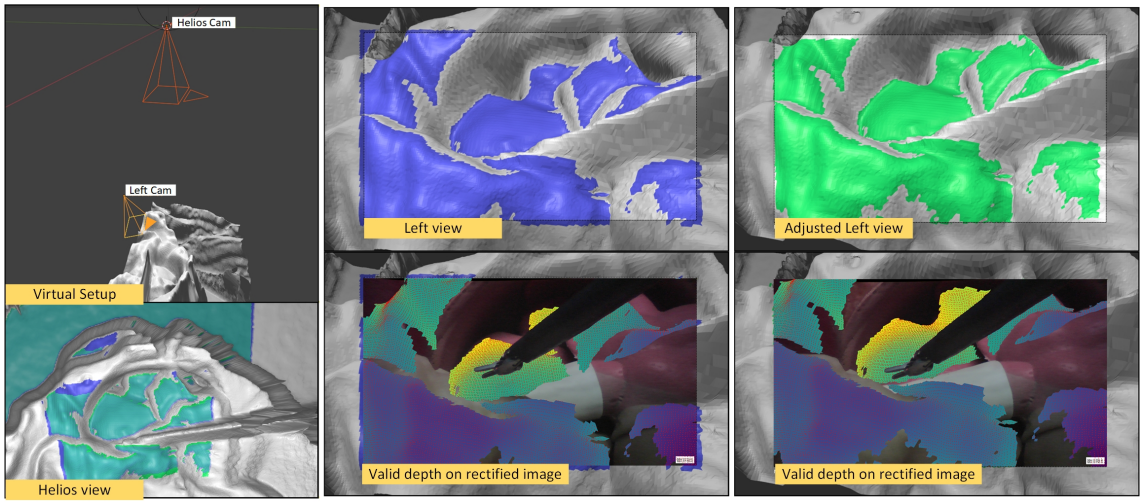


Figure 5.4: Transformation pipeline. The virtual setup set the cameras' positions according to the calibration results (first column). The left view shows the results of re-projecting the mesh onto the left image plane (second column). After manually adjusting the left camera position within the error margin, the transformation to move the camera is stored and used to produce new re-projections (third column). Notice that the left camera views of the virtual environment show the pixels with valid depth information before and after the adjustment, blue and green, respectively.

rendered and overlaid on the RGB information (second column in Figure 5.4). Therefore, to account for residual misalignment within the transformation error margin (error defined by the mean difference between the triangulated and transformed 3D points $e = \frac{1}{N} \sum ||\mathbf{X}_i^{\text{RGB}} - \hat{\mathbf{X}}_i^{\text{RGB}}||^2$), we defined a compensation transformation. This was obtained by replicating the setup in a 3D modelling environment (Blender) and manually adjusting the endoscopic camera position until the rendered and real RGB views were aligned (third column in Figure 5.4). The resulting offset defines the compensation transformation, which is shared within the dataset.

Ex Vivo image denoising

The endoscopic images obtained during ex vivo data collection exhibit fine-grained, spatially uniform speckle noise. To denoise the images, we share the code to apply patch-based non-local means filtering [153] with a search window of 21×21 pixels and a luminance strength weight of 8 to reduce luminance and chrominance noise while preserving fine tissue structure. Figure 5.5 shows an example of an image before and after the denoiser. Our dataset contains raw, noisy images, and the code in our repository includes the weighting, modification, and application of our denoising function.



Figure 5.5: Denoising Ex-Vivo image

5.3 Data Records

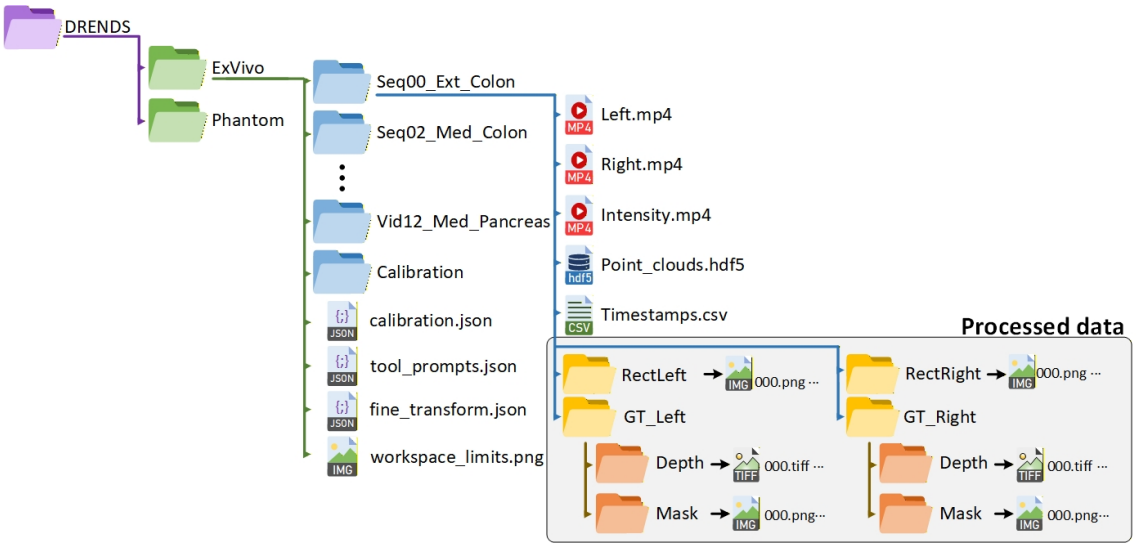


Figure 5.6: Data folder structure in DRENDS

The DRENDS dataset contains synchronised stereo endoscopic images, point clouds, and metadata, together with the code to generate binary masks and depth

maps. Figure 5.6 illustrates the folder hierarchy and file formats. Data folders are named according to organ, illumination condition, and acquisition mode (sequence or video). Each recording is stored in a dedicated folder that includes the raw data: videos, point clouds, and timestamps. Once the processing pipeline is run, rectified images, masks, and depth maps are produced. File formats are consistent across the dataset: videos are .mp4, images are .png, depth maps are .tiff, point clouds are .hdf5, and timestamps are .csv. All data are stored in 16bit format.

The dataset comprises 54 folders, corresponding to two modalities (phantom and ex vivo), five abdominal organs (colon, intestine, stomach, liver, and pancreas - liver is excluded in the ex vivo data), three illumination conditions (high, standard, and external) and two acquisition approaches (sequences and videos). In total, the dataset contains 34,650 frames with depth ground truth. Figure 5.7 shows a representative example of the data. It displays the stereo images, associated metadata and the corresponding processed depth maps and masks. This example illustrates the type and quality of the information available for each time step in the sequences.

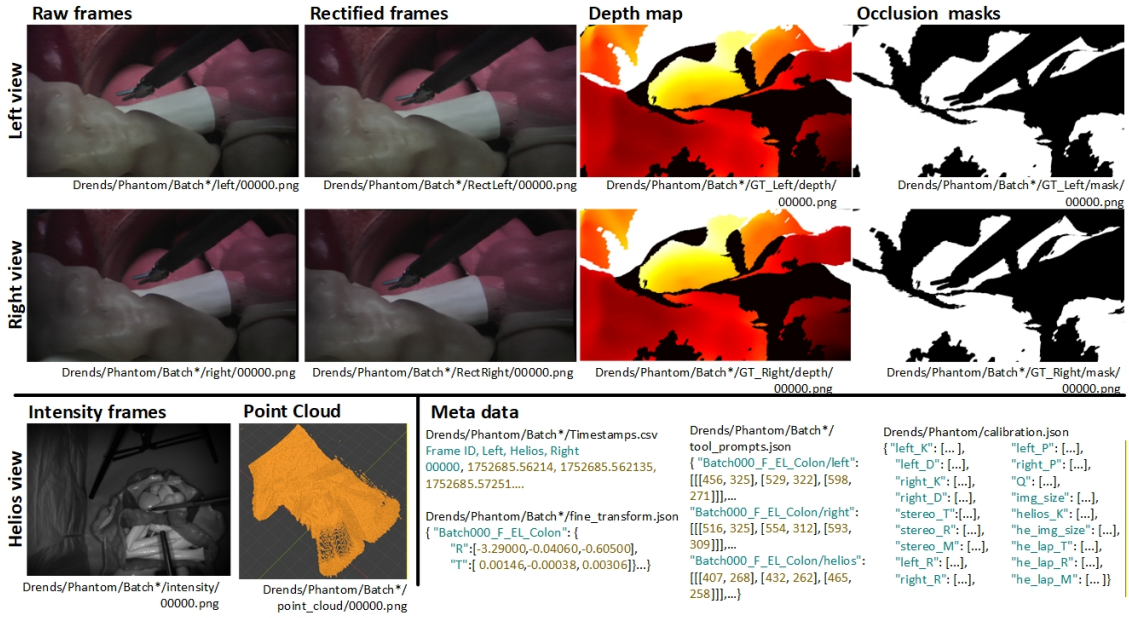


Figure 5.7: Example of the data related to a single time step in a sequence.

5.4 Technical Validation

5.4.1 Calibration

A full stereo calibration was performed for each setup to ensure accurate geometric reconstruction between the left and right cameras using checkerboard images. The resulting parameters were used to triangulate the pattern's points. The 3D information from the ToF camera is then aligned to the triangulated pattern's points. Table 5.1 shows the results of the calibration and the 3D alignments. The calibration results are used to initialise the virtual environment and obtain a fine transformation that aligns the 3D information with the stereo camera view.

Table 5.1: Camera calibration results

Type	**Image resolution**	**Camera calibration**			**ToF-Stereo transformation**	
		left	*right*	*stereo*	*3D alignment error*	*2D projection error*
Phantom	1280×720px	0.424 px	0.415 px	0.504 px	1.43 +/- 0.93 mm	6.28 +/- 4.00 px
Ex vivo	1920×1080px	0.332 px	0.353 px	0.523 px	4.25 +/- 2.21 mm	8.73 +/- 5.73 px

The temporal synchronisation accuracy was quantified by measuring inter-frame timestamp differences between sensing modalities. Across the phantom videos, the average time offset between stereo pairs was 8 ms (maximum 50 ms), indicating tight synchronisation between the stereo cameras. The results on the RGB–ToF alignment reveal a capture delay in one of the videos. The video 'Vid11_Liver_High' contains 16 frames with a time offset ranging from 280 ms to 1,792 ms. Discarding this outlier, the average time offset is 33 ms and the maximum time offset is 131 ms. Considering the target time offset of 62 ms for an acquisition speed of 16 FPS, only 2.2 timesteps per video have a time offset higher than the target time.

Across the ex vivo videos, the average time offset between stereo pairs was 9 ms (maximum 72 ms), with only a single instance exceeding the nominal frame interval. The RGB–ToF alignment shows consistency, with an average offset of 12 ms and a maximum of 92 ms. On average, 2.2 time steps per video have a time offset between sensors exceeding 62 ms. Overall, the temporal offsets remain small relative to the acquisition rate for the vast majority of frames, ensuring reliable temporal alignment for dynamic scene analysis.

5.4.2 Benchmark of state-of-the-art methods

To evaluate the performance of several (SOTA) depth estimation methods, we selected well-established models and applied them to the image sequences in our dataset. All models were used in their publicly released form, without additional

fine-tuning or dataset-specific adaptation, to provide a fair "out-of-the-box" comparison. The evaluation focuses on both spatial accuracy and temporal consistency, using metrics described in the following subsection.

Evaluation Metrics

To quantitatively assess the performance of depth estimation models on DRENDS, we employ a set of metrics that evaluate both spatial accuracy and temporal consistency. The metrics are computed at the pixel level over valid regions (excluding occluded pixels using the masks) and reported as averages.

Spatial accuracy metrics. For each frame, we denote the ground-truth depth as d_i and the predicted depth as $\hat{d}_i$, where i indexes the valid pixels. Following common practice in depth estimation benchmarks, we report the following metrics:

- **Mean Absolute Error (MAE)** in millimetres (mm):

$$\text{MAE} = \frac{1}{N}\sum_i |d_i - \hat{d}_i| \tag{5.2}$$

- **Root Mean Squared Error (RMSE)** in millimetres (mm):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_i (d_i - \hat{d}_i)^2} \tag{5.3}$$

- **Log RMSE**:

$$\text{RMSE}_{\log} = \sqrt{\frac{1}{N}\sum_i \left(\log d_i - \log \hat{d}_i\right)^2} \tag{5.4}$$

- **Absolute Relative Error (AbsRel)**:

$$\text{AbsRel} = \frac{1}{N}\sum_i \frac{|d_i - \hat{d}_i|}{d_i} \tag{5.5}$$

- **Threshold accuracy**: proportion of pixels for which the difference between predicted and ground-truth depth lies within a tolerance δ in mm.:

$$\delta_{Threshold} = \frac{1}{N}\sum_i \left(\max\left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}\right) < \delta^k\right), \qquad \delta = 1.25,\ k \in \{0.25, 0.50, 0.75\}. \tag{5.6}$$

These metrics jointly measure absolute and relative discrepancies between predicted and true depths. While MAE and RMSE quantify absolute deviations in mil-

limetres (mm), $\text{RMSE}_{\log}$ and AbsRel account for proportional and scale-dependent errors. The δ-accuracy terms indicate the fraction of pixels within different error tolerances.

Aligned-depth evaluation. Because depth estimation models may produce outputs that are accurate in structure but biased in global scale or offset, we additionally report the above metrics after rescaling predictions to the ground truth. The aligned depth $\hat{d}'_i$ is obtained by minimising the least-squares difference over all valid (non-occluded) pixels in a sequence:

$$\min_{s,t} \sum_i (s\,\hat{d}_i + t - d_i)^2, \qquad \hat{d}'_i = s\,\hat{d}_i + t, \tag{5.7}$$

where s and t denote the scale and translation parameters, respectively. The parameters are estimated once per sequence and applied consistently to all frames within that sequence, rather than being computed independently for each frame. This choice avoids compensating frame-to-frame prediction fluctuations through independent rescaling and therefore preserves temporal consistency during evaluation. This evaluation highlights the structural fidelity of the predictions independently of global scaling errors.

Temporal consistency metric. While per-frame metrics such as MAE and RMSE quantify absolute depth accuracy, they do not capture the temporal stability of predictions across frames. For example, two models may achieve similar per-frame errors, yet one may produce temporally stable depth estimates while the other exhibits frame-to-frame fluctuations (flickering) due to noise or inconsistent scale estimation. In such cases, the latter would exhibit poorer temporal consistency despite comparable per-frame accuracy. Recent work in surgical vision has further shown that enforcing temporal consistency can significantly improve depth prediction performance in endoscopic settings [154]. Therefore, we propose a metric that explicitly captures this behaviour by evaluating the agreement between predicted and ground-truth depth changes over time, which can be used for both evaluation and training of depth estimation methods. Let $d_i^{(t)}$ and $\hat{d}_i^{(t)}$ denote the ground-truth and predicted depths for frame t. We define:

- **Depth Temporal Consistency Error (DTCE)**: This metric assesses how accurately the predicted depth differences between frames align with the true temporal depth changes, i.e., the average per-pixel jitter of the prediction, in mm. Lower values indicate better temporal consistency, as they correspond to the average absolute difference between the predicted and ground-truth depth

variations across consecutive frames.

$$\mathrm{DTCE} = \frac{1}{N} \sum_{i} \left| (\hat{d}_i^{(t+1)} - \hat{d}_i^{(t)}) - (d_i^{(t+1)} - d_i^{(t)}) \right| \tag{5.8}$$

Collectively, this temporal metric complements the frame-wise spatial evaluation by assessing the stability and coherence of predicted depth sequences. It provides a quantitative basis for evaluating models on the dynamic, deformable scenarios characteristic of robotic-assisted laparoscopic surgery.

Models

Numerous approaches have adapted and fine-tuned general-purpose depth estimation models to the surgical domain, with notable gains in in-distribution accuracy [155]. The objective of the present benchmark is distinct and complementary: we evaluate the publicly released, off-the-shelf foundation models from which many of these surgical-specific approaches are derived or have been used as baselines, and we then evaluate the surgery-adapted models themselves, so that the two can be read against one another on the same sequences and under the same protocol. Reporting their out-of-the-box performance on DRENDS establishes a common, reproducible reference point against which any domain-adapted method can be positioned. To this end, we benchmark thirteen publicly released depth estimation models, grouped into four classes that differ fundamentally in what their outputs represent and, consequently, in the alignment required to compare them against ground truth: relative monocular, metric monocular, metric stereo, and models tailored for surgical data. The last class collects recent surgery-specific depth models, so that the comparison is made against the current state of the art in the domain.

Relative monocular depth **Monodepth2** [115], **MiDaS** [156], and the relative variant of **Depth Anything Model V2** [157] (**DAMv2**) are trained with a scale-and-shift-invariant (SSI) objective in inverse-depth space. The SSI loss, introduced for MiDaS, computes a closed-form per-image scale and shift that best aligns the network output to the ground-truth inverse depth on the valid pixels, and subtracts that fit out before scoring the residual; the network is therefore explicitly trained to discard the scale and shift of its inverse-depth output. Monodepth2 is self-supervised on monocular video via photometric reprojection and is scale-ambiguous only, since the reprojection loss fixes the zero point of disparity; MiDaS and Depth Anything V2 are trained on heterogeneous mixtures of depth datasets and are ambiguous up to both scale and shift. The output of all three models is therefore affine-related

to the ground-truth inverse depth, not to depth itself. It is important to notice that we accordingly adopt the inverse-depth alignment under which these models were originally trained: we solve Equation 5.7 after replacing d and $\hat{d}$ with $1/d$ and $1/\hat{d}$, fit $(s^\star, t^\star)$ in inverse-depth space, apply the linear map there, and invert the result back to depth before computing metrics. Because the raw output of these models is in arbitrary inverse-depth units, metric error (MAE, RMSE, AbsRel, δ_k) is undefined before alignment and is reported only for the aligned variant.

Metric monocular depth **ZoeDepth** [158], **Depth Pro** [155], and the metric variant of **Depth Anything V2** [157] (DAMv2-metric) predict depth in absolute units (metres). ZoeDepth couples a MiDaS-style SSI backbone with metric bin heads that express depth as a weighted combination of learnable bin centres; the relative pretraining provides cross-domain generalisation while the bin heads, fine-tuned on labelled metric data, anchor the output to a physical scale and route between domains at inference time. Depth Pro is a multi-scale vision transformer trained with large-scale supervised metric depth on a curated dataset mixture, designed to produce sharp object boundaries and to remain metric without external intrinsics: a focal-length head is trained jointly and supplies the camera parameters at inference when none are provided. DAMv2-metric reuses the self-distilled relative DAMv2 backbone and fine-tunes it for metric prediction on synthetic indoor RGB-D data. All three were trained outside the endoscopic domain, so a residual systematic bias on DRENDS is expected; we therefore report the raw predictions, which represent the deployable accuracy a user obtains by running the off-the-shelf model, and additionally an aligned variant using Equation 5.7 in which a per-sequence linear fit absorbs that residual bias. The gap between raw and aligned metrics quantifies the domain shift between each model's training distribution and DRENDS.

Metric stereo depth **RAFT-Stereo** [159], **UniMatch** [160], and **FoundationStereo** [161] predict per-pixel horizontal disparity δ between the rectified stereo pair. RAFT-Stereo iteratively refines disparity on a multi-level correlation volume using a recurrent update operator. UniMatch treats stereo matching as a dense correspondence problem within a unified transformer architecture shared with optical flow and depth estimation, using cross-attention between left and right feature maps to compute disparity in a single forward pass. FoundationStereo is a transformer-based zero-shot stereo model that builds on a vision foundation backbone and is trained on a large mixture of synthetic and real stereo data, generalising without per-domain fine-tuning. Disparity is converted to metric depth using the standard

pinhole relation:

$$\hat{d}_i \;=\; \frac{f \cdot B}{\delta_i}, \qquad f \;=\; P_{\text{left}}[0,0], \qquad B \;=\; -\,\frac{P_{\text{right}}[0,3]}{f} \cdot 10^3, \tag{5.9}$$

where f is the rectified focal length in pixels, and B is the rectified baseline in millimetres. Both are calculated per sequence from the rectified projection matrices $P_{\text{left}}, P_{\text{right}} \in \mathbb{R}^{3\times4}$ that are provided in the calibration files. The use of rectified projection matrices (rather than the raw intrinsics K) is essential: the rectified focal length differs from the unrectified focal length by a homography-dependent factor, and interchanging them introduces a silent bias to every predicted depth by a multiplicative constant. As with the metric monocular models, we report both the raw stereo predictions and a per-sequence depth-space aligned variant to make any residual calibration bias explicit.

Models tailored for surgical data **DARES** [162], **EndoDAC** [163], **Endo-FASt3r** [164] and **EndoUFM** [165] adapt the same foundation backbones evaluated above to endoscopic imagery, and all four are trained self-supervised on surgical video through photometric reprojection rather than on depth labels. DARES adapts Depth Anything V2 through a Vector-LoRA scheme that assigns higher ranks to the early layers of the encoder and lower ranks to the later ones; EndoDAC adapts the same backbone through a dynamic vector-based low-rank adaptation together with convolutional neck blocks and a pose-intrinsics network that predicts the camera parameters, so that no calibration is required at training time; Endo-FASt3r pairs a Depth Anything V2 depth branch with a Reloc3r-derived pose branch to solve depth and ego-motion jointly as a structure-from-motion problem; and EndoUFM combines the Depth Anything backbone with MedSAM segmentation, adapting it through a random-vector low-rank scheme and a mask-guided smoothness loss. Because all four inherit the scale ambiguity of self-supervised monocular training, and are evaluated in their original works after median scaling against the ground truth, their outputs are affine-related to depth in the same sense as the relative monocular models, and we therefore report them under the aligned protocol of Equation 5.7 alone.

Two further surgery-specific methods were considered and could not be included in the comparison. **Surgical-DINO** [166] adapts DINOv2 for supervised depth estimation in endoscopy through LoRA layers inserted into a frozen image encoder, and **StereoMamba** [167] estimates stereo disparity through a feature-extraction Mamba module designed to capture long-range spatial dependencies within and across the stereo pair. Neither releases weights trained on surgical data, so they are

described here for completeness and excluded from the tables.

Results

We evaluate the thirteen models on both DRENDS subsets for raw and aligned predictions, and report spatial accuracy (MAE, RMSE, RMSE-log, AbsRel), threshold accuracy (δ), and temporal consistency (DTCE), with each metric computed across the three illumination conditions. Tables 5.2 and 5.3 report the Phantom and ex vivo results, Figures 5.8 and 5.9 present the corresponding per-illumination breakdown, and Figures 5.10 and 5.11 present qualitative results of the models of one image per anatomy. The models tailored for surgical data are reported separately in Table 5.4 and Figure 5.12, since they are evaluated under the aligned protocol alone.

The metrics for the raw predictions cleanly separate the models by the source of their absolute scale. Stereo models are the closest to out-of-the-box relevance, achieving approximately 18 mm MAE on Phantom (Table 5.2) and 12 mm on ex vivo (Table 5.3) (AbsRel 0.10 and 0.13, respectively). On the other hand, the metric monocular models infer absolute scale from priors tied to their training distribution, which is far out of distribution for laparoscopic imagery. That learned scale produces errors of hundreds of millimetres, up to $\sim$850 mm, with AbsRel between 2 and 7, and essentially no pixels within the tolerance bands ($\delta \approx 0$). The relative monocular models carry no absolute scale by construction and are undefined in the metric space; they are reported only after alignment.

Per-sequence alignment brings the prediction depth range into the GT range (Figures 5.10 and 5.11) and exposes the geometric quality of each prediction. All models converge to low-millimetre error, confirming that the large raw gaps were a matter of shift and scale rather than structure. The stereo methods remain the most accurate, clustering near 7–8 mm MAE on Phantom and 5 mm on ex vivo, with the highest threshold accuracy. Among the monocular models, the relative variant of DAMv2 is the strongest (10.48 mm on Phantom, 6.75 mm on ex vivo), narrowly ahead of MiDaS and of its own metric counterpart. In general, off-the-shelf models recover scene geometry competently but in an out-of-distribution depth range.

Temporal consistency matters independently of absolute accuracy in surgical video, where frame-to-frame jitter undermines tracking, registration, and 3D reconstruction. The large raw DTCE of DepthPro and DAMv2-metric is a combination of jitter and scale error; inter-frame motion is amplified, and it decreases once alignment restores the scale. Among the stable models, stereo methods stand out with the lowest DTCE under the raw protocol and remain among the most stable after

Table 5.2: **Comparison of depth estimation SOTA methods across various metrics on DRENDS-Phantom sequences**

Model	MAE (↓) (in *mm*)	RMSE (↓) (in *mm*)	RMSE-log (↓)	AbsRel (↓)	Depth Accuracy (↑) $\delta^{0.25}$	$\delta^{0.50}$	$\delta^{0.75}$	DTCE (↓) (in *mm*)
Non aligned (RAW)								
ZoeDepth	849.39±25.60	849.98±25.01	2.00±0.11	6.47±0.86	0.00±0.00	0.00±0.00	0.00±0.00	13.34±10.21
DepthPro	276.62±67.66	286.23±67.17	1.06±0.17	2.05±0.54	0.00±0.01	0.01±0.01	0.01±0.02	102.20±30.18
DAMv2-Metric	378.38±52.12	388.44±53.25	1.32±0.12	2.81±0.48	0.00±0.00	0.00±0.00	0.00±0.00	55.82±9.27
FoundationStereo	17.86±4.33	21.09±4.61	0.16±0.03	0.13±0.02	0.10±0.09	0.41±0.14	0.77±0.10	6.02±2.01
RAFT-Stereo	17.98±4.53	21.35±4.78	0.17±0.03	0.13±0.02	0.12±0.08	0.41±0.15	0.75±0.12	6.83±2.27
UniMatch	18.00±4.66	21.23±4.81	0.16±0.03	0.13±0.02	0.11±0.09	0.41±0.15	0.76±0.12	6.03±1.99
Aligned (ALGN)								
Monodepth2	15.79±3.18	19.94±3.57	0.14±0.02	0.11±0.02	0.30±0.11	0.57±0.11	0.78±0.09	5.59±1.87
MiDaS	12.02±2.12	15.62±2.75	0.11±0.02	0.09±0.02	0.42±0.09	0.70±0.09	0.88±0.05	7.97±2.44
DAMv2	10.48±2.17	13.73±2.83	0.10±0.02	0.08±0.01	0.48±0.07	0.77±0.07	0.91±0.04	7.72±2.36
ZoeDepth	15.34±2.85	19.09±3.18	0.14±0.02	0.11±0.02	0.30±0.11	0.56±0.13	0.80±0.08	5.93±2.21
DepthPro	14.27±2.58	17.75±2.80	0.13±0.02	0.11±0.02	0.33±0.08	0.61±0.11	0.82±0.07	8.87±2.46
DAMv2-Metric	11.75±2.34	14.94±2.79	0.11±0.02	0.09±0.02	0.41±0.08	0.71±0.09	0.89±0.06	8.00±1.99
FoundationStereo	**7.24±1.43**	**10.76±2.19**	**0.08±0.01**	**0.05±0.01**	**0.68±0.08**	**0.91±0.03**	**0.95±0.02**	**5.32±1.96**
RAFT-Stereo	7.82±1.53	11.43±2.07	**0.08±0.01**	0.06±0.01	0.64±0.09	0.88±0.04	0.94±0.02	6.00±2.16
UniMatch	7.43±1.48	10.91±2.10	**0.08±0.01**	**0.05±0.01**	0.66±0.08	0.90±0.04	**0.95±0.02**	5.48±2.03

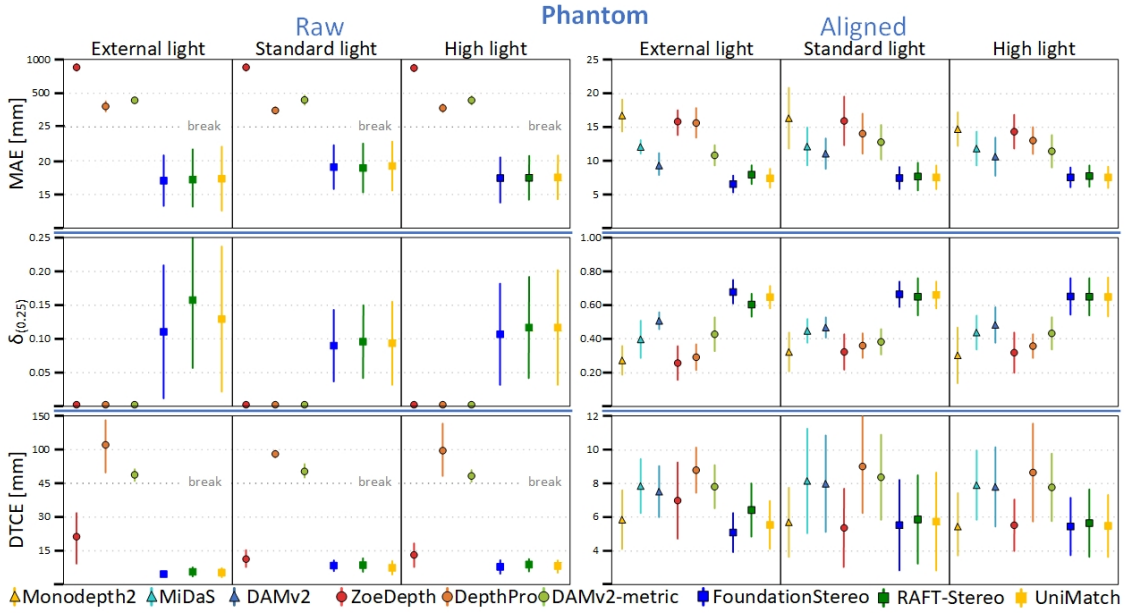


Figure 5.8: **Benchmarking over DRENDS-Phantom** under all illumination (*External Light*, *Standard Light*, and *High Light*). Results from all different sequences are combined and reported per model, illumination and output type (raw and aligned).

alignment. ZoeDepth is an interesting case: despite having the worst spatial accuracy, its raw DTCE is among the lowest (13.3 mm on Phantom, 3.3 mm on ex vivo), because its scale error stays nearly constant across frames, showing that temporal smoothness and absolute accuracy are distinct properties.

The per-illumination breakdown shows that lighting is not a primary source of error. Within each family, the metrics vary only modestly across the external, standard, and high-illumination conditions (Figures 5.8 and 5.9), and this variation is

Table 5.3: **Comparison of depth estimation SOTA methods across various metrics on DRENDS-Ex vivo sequences**

Model	MAE (↓) (in *mm*)	RMSE (↓) (in *mm*)	RMSE-log (↓)	AbsRel (↓)	Depth Accuracy (↑) $\delta^{0.25}$	$\delta^{0.50}$	$\delta^{0.75}$	DTCE (↓) (in *mm*)
Non aligned (RAW)								
ZoeDepth	870.28±3.76	870.35±3.77	2.05±0.03	6.79±0.25	0.00±0.00	0.00±0.00	0.00±0.00	**3.32±1.16**
DepthPro	572.88±174.42	575.93±173.78	1.63±0.27	4.46±1.41	0.00±0.00	0.00±0.00	0.00±0.00	121.72±38.07
DAMv2-Metric	597.51±80.97	602.39±81.98	1.71±0.13	4.64±0.68	0.00±0.00	0.00±0.00	0.00±0.00	84.58±11.62
FoundationStereo	12.44±3.00	14.18±2.80	0.12±0.03	0.10±0.02	0.22±0.17	0.63±0.19	0.92±0.04	4.37±1.15
RAFT-Stereo	12.46±3.00	14.17±2.79	0.12±0.03	0.10±0.02	0.22±0.17	0.62±0.19	0.92±0.04	4.50±1.16
UniMatch	12.51±2.94	14.18±2.66	0.12±0.02	0.10±0.02	0.22±0.16	0.62±0.19	0.92±0.04	4.22±1.10
Aligned (ALGN)								
Monodepth2	8.28±2.00	10.43±2.56	0.08±0.02	0.06±0.02	0.53±0.11	0.85±0.09	0.96±0.04	3.88±1.09
MiDaS	7.30±1.56	9.11±1.90	0.07±0.02	0.06±0.01	0.59±0.11	0.88±0.07	0.98±0.02	4.66±1.79
DAMv2	6.75±0.92	8.47±1.20	0.07±0.01	0.05±0.01	0.62±0.07	0.91±0.05	0.98±0.02	4.51±1.59
ZoeDepth	8.76±1.81	10.85±2.31	0.08±0.02	0.07±0.02	0.49±0.09	0.84±0.08	0.95±0.04	3.07±1.16
DepthPro	8.57±1.76	10.65±2.28	0.08±0.02	0.07±0.02	0.50±0.09	0.84±0.08	0.96±0.04	3.67±1.26
DAMv2-Metric	7.92±1.53	9.89±1.95	0.08±0.02	0.06±0.01	0.55±0.09	0.86±0.07	0.96±0.03	4.35±1.49
FoundationStereo	4.86±0.96	6.56±1.26	**0.05±0.01**	**0.04±0.01**	0.80±0.09	**0.96±0.03**	**0.98±0.01**	3.44±1.02
RAFT-Stereo	4.85±0.95	6.56±1.24	**0.05±0.01**	**0.04±0.01**	0.80±0.08	**0.96±0.03**	**0.98±0.01**	3.53±0.98
UniMatch	**4.78±0.99**	**6.47±1.28**	**0.05±0.01**	**0.04±0.01**	**0.81±0.09**	**0.96±0.03**	**0.98±0.01**	3.42±1.02

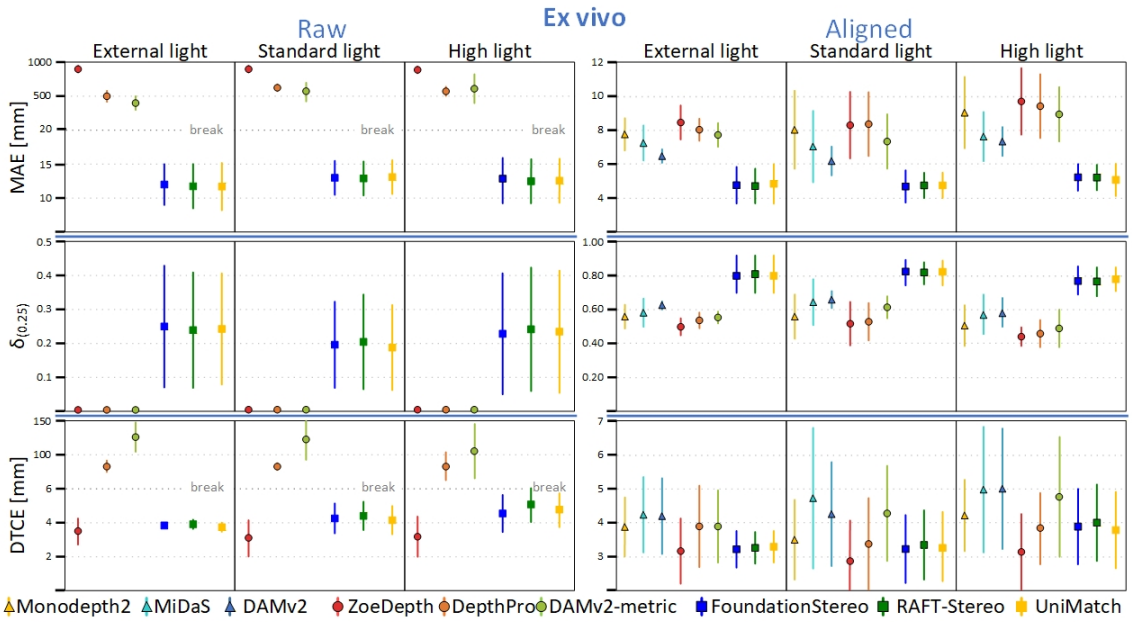


Figure 5.9: **Benchmarking over DRENDS-ExVivo** under all illumination (*External Light*, *Standard Light*, and *High Light*). Results from all different sequences are combined and reported per model, illumination and output type (raw and aligned).

small relative to the differences between models and between subsets. The dominant limitations are therefore the domain shift and, for the monocular models, the absolute scale, rather than the photometric conditions under which the sequences were captured. Figures 5.10 and 5.11 present qualitative depth maps on representative phantom and ex vivo frames. Most methods recover the scene's geometry at the wrong scale, as evidenced by their low alignment errors and depth ranges, but differ in their finer structure. The discrepancies concentrate in the regions with low-

textured, homogeneous surfaces and strong specular highlights, where photometric cues are weak.

The models tailored for surgical data behave, on these sequences, like the relative monocular family from which they are derived. Under the raw protocol their predictions are expressed in arbitrary inverse-depth units, which places the MAE at approximately 135 mm on Phantom and 129 mm on ex vivo with an AbsRel of 1.00 and no pixels inside any tolerance band, so only the aligned metrics of Table 5.4 carry information. Once aligned, the four models cluster tightly and sit in the middle of the benchmark: on ex vivo they span 7.35 to 7.56 mm MAE, behind the stereo methods at 4.78–4.86 mm and behind DAMv2 at 6.75 mm, and marginally behind MiDaS at 7.30 mm, while ahead of DAMv2-Metric, Monodepth2, DepthPro and ZoeDepth; on Phantom they span 12.22 to 13.14 mm, behind stereo at 7.24–7.82 mm and behind both variants of DAMv2, and comparable to MiDaS at 12.02 mm. Their temporal consistency follows the same pattern, with DTCE between 3.73 and 3.88 mm on ex vivo and between 6.32 and 6.40 mm on Phantom, close to the stereo methods and ahead of the relative monocular models on Phantom but behind ZoeDepth on ex vivo. The ordering between the four is narrow enough, relative to the standard deviations across sequences, that we do not read it as separating them: EndoDAC is nominally first on Phantom and DARES on ex vivo, but the differences are a fraction of the spread of either.

Figure 5.12 shows the corresponding qualitative behaviour. The predicted structure is coherent and the tool is separated from the tissue in every case, but the raw ranges reveal how far the outputs sit from a metric quantity: most frames are predicted within a span of 0.1 to 0.5 mm, while on individual sequences the same model returns ranges of hundreds or thousands of millimetres, so the units are neither metric nor stable across sequences. After alignment the ranges agree with the ground truth to within a few millimetres, which confirms that what these models recover is the geometry of the scene rather than its scale.

Benchmark conclusion

Our benchmark establishes a reproducible reference point for off-the-shelf depth estimation in robotic endoscopy and yields a clear practical verdict: stereo methods are the closest to being deployable out of the box, recovering depth via triangulation from the known rectified camera parameters; they provide a geometric reference for the surgical scenario without any learned scale. It is important to note that stereo methods still exhibit residual error, reflecting the difficulty of stereo correspondence in the surgical setting, where low-textured, homogeneous tissue and strong specular

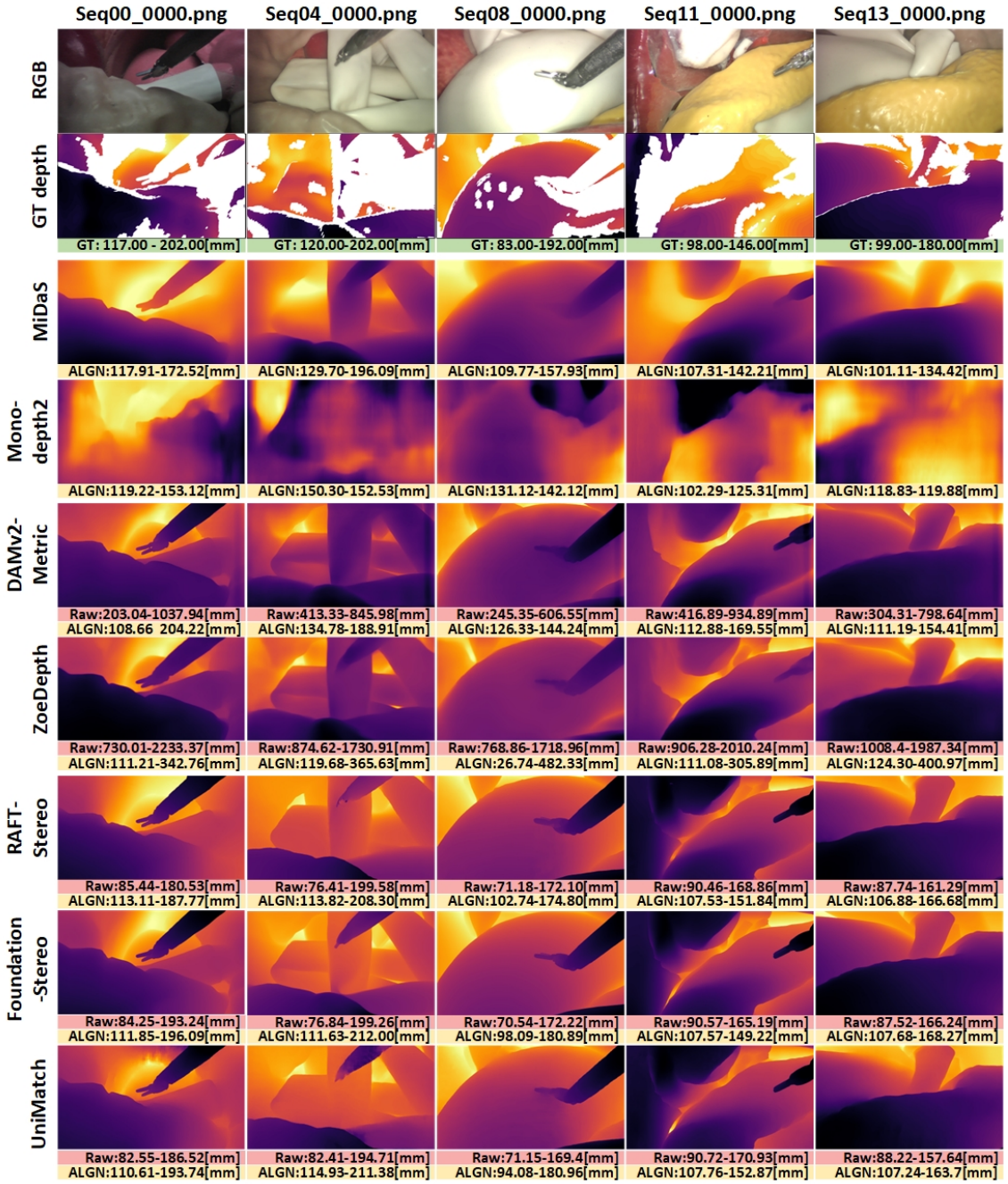


Figure 5.10: Qualitative depth estimation comparison of SOTA models over an image frame per organ in the DRENDS phantom subset. The depth range for ground truth (GT), unaligned (RAW), and aligned (ALGN) is presented for each frame.

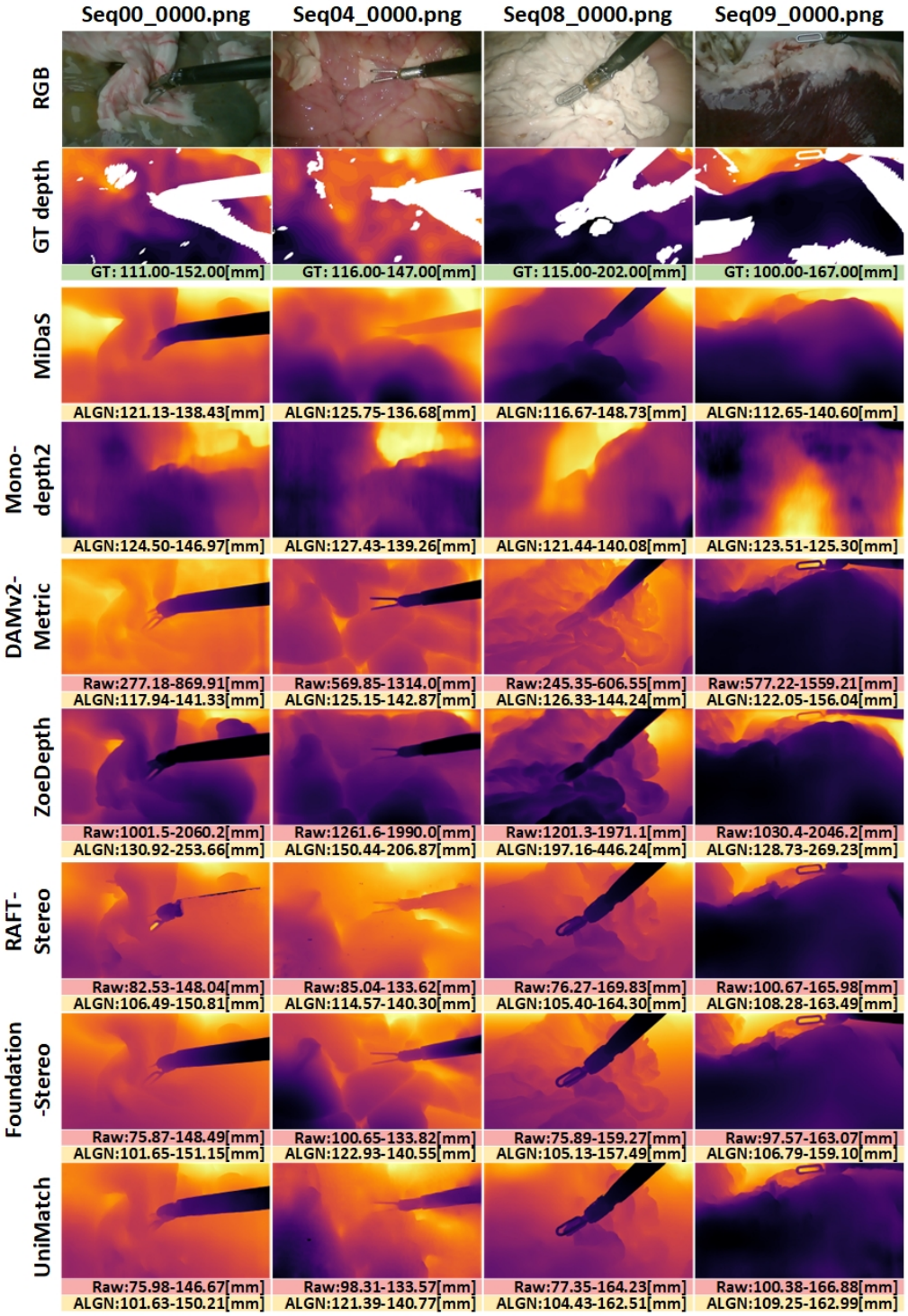


Figure 5.11: Qualitative depth estimation comparison of SOTA models over an image frame per organ in the DRENDS ex vivo subset. The depth range for ground truth (GT), unaligned (RAW), and aligned (ALGN) is presented for each frame.

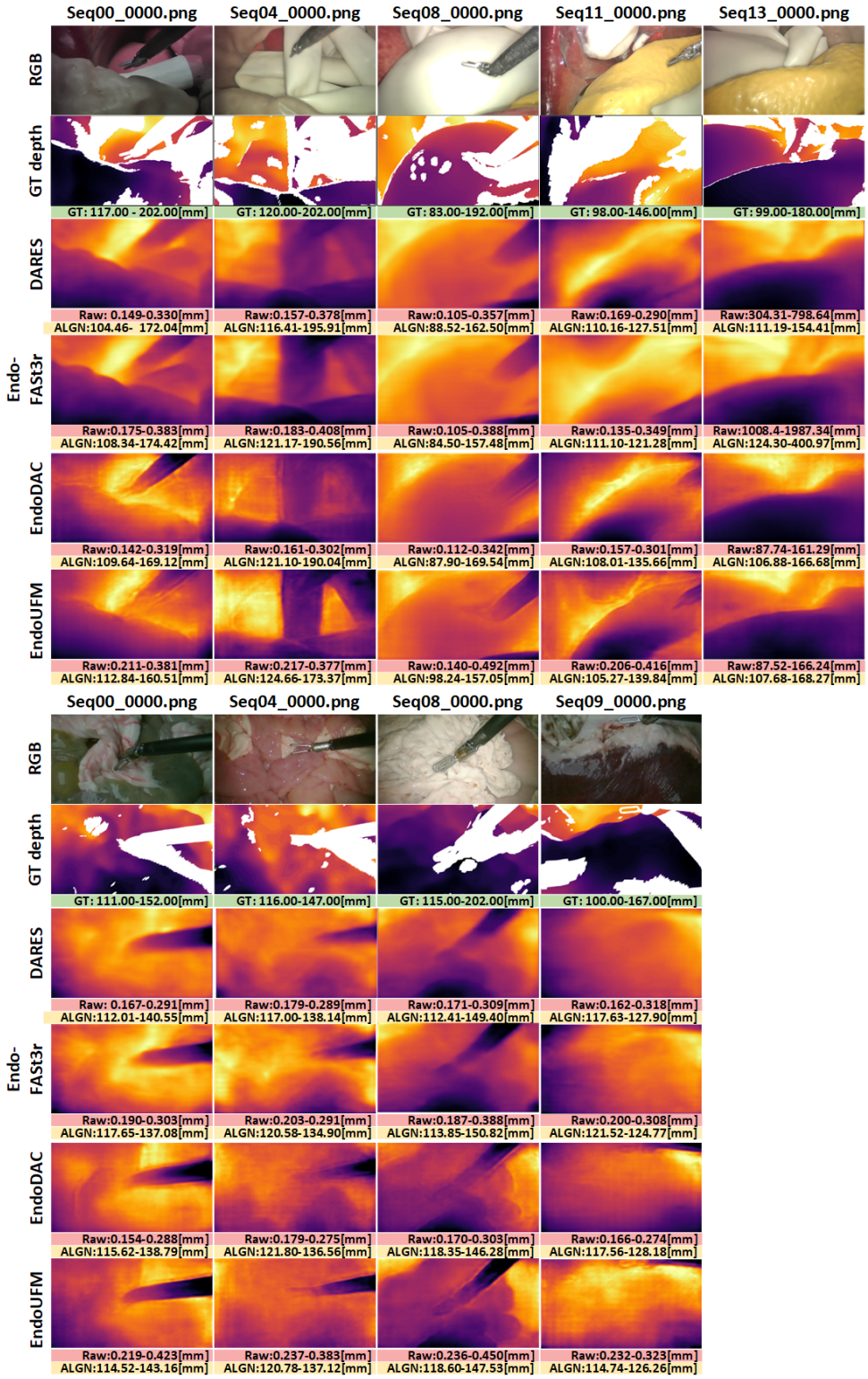


Figure 5.12: Qualitative depth estimation comparison of the models tailored for surgical data over an image frame per organ, on the DRENDS phantom subset (top) and the ex vivo subset (bottom). The depth range for ground truth (GT), unaligned (RAW) and aligned (ALGN) is presented for each frame.

Table 5.4: **Comparison of depth estimation models tailored for surgical data across various metrics on both DRENDS subsets. All four models are scale-ambiguous, so only the aligned protocol is reported**

Model	MAE (↓) (in *mm*)	RMSE (↓) (in *mm*)	RMSE-log (↓)	AbsRel (↓)	Depth Accuracy (↑) $\delta^{0.25}$	$\delta^{0.50}$	$\delta^{0.75}$	DTCE (↓) (in *mm*)
DRENDS-Phantom, aligned (ALGN)								
DARES	12.46±2.35	15.85±2.55	0.11±0.02	0.09±0.02	0.38±0.12	0.68±0.14	0.87±0.12	6.33±2.52
EndoDAC	**12.22±2.34**	**15.58±2.71**	0.11±0.02	0.09±0.02	**0.39±0.12**	**0.69±0.13**	**0.88±0.10**	6.40±2.55
Endo-FASt3r	12.86±2.24	16.50±2.71	0.12±0.02	0.09±0.02	0.37±0.10	0.66±0.11	0.87±0.07	**6.32±2.50**
EndoUFM	13.14±2.53	16.73±2.94	0.12±0.02	0.10±0.02	0.37±0.12	0.64±0.13	0.85±0.10	6.37±2.63
DRENDS-ExVivo, aligned (ALGN)								
DARES	**7.35±1.74**	**9.28±2.13**	0.07±0.02	0.06±0.01	**0.59±0.12**	0.88±0.08	0.97±0.03	3.88±1.58
EndoDAC	7.44±1.54	9.33±1.94	0.07±0.02	0.06±0.01	0.58±0.10	0.88±0.07	0.97±0.03	**3.73±1.47**
Endo-FASt3r	7.56±1.66	9.54±2.04	0.07±0.02	0.06±0.01	0.58±0.11	0.88±0.08	0.97±0.03	3.74±1.49
EndoUFM	7.45±1.41	9.33±1.78	0.07±0.01	0.06±0.01	0.57±0.10	0.88±0.07	0.97±0.02	3.77±1.61

highlights diminish the photometric cues on which matching depends. Monocular models cannot produce metric depth out of the box, despite recovering the 3D shape of the scenario. Their gap then is not only the photometric challenges but a missing metric anchor, which suggests that adapting monocular models to surgery should focus on recovering scale through sparse metric cues, known camera geometry, or a lightweight calibration step rather than on re-learning the geometry they already capture.

Temporal consistency emerges as an axis of performance distinct from spatial accuracy, and any benchmark for surgical video should report it alongside spatial accuracy. ZoeDepth illustrates the decoupling most clearly: it is the least accurate model in absolute terms yet among the most temporally stable, because its scale error stays nearly constant across frames and therefore introduces little frame-to-frame jitter. A model can thus be poorly scaled yet smooth in time, or accurate yet unstable, so neither property can be inferred from the other. Here, temporal stability refers to a smooth and proportional response to camera movements and deformations (rigid and non-rigid). In downstream tasks such as tracking, registration, and 3D reconstruction, where stability matters as much as per-frame accuracy, both must be measured explicitly.

Finally, the two subsets isolate distinct facets of the problem. Although Phantom is synthetic, it is consistently the harder regime: across all models, the aligned errors are higher, the threshold accuracy is lower, and the temporal consistency is worse than on ex vivo. This follows from its wider depth range and its smooth, low-texture, specular surfaces, which magnify scale-sensitive errors and destabilise temporal predictions, whereas the ex vivo sequences, despite the greater visual complexity of real tissue and non-rigid deformation, span a narrower depth range and are correspondingly easier once scale is corrected. The dominant difficulty in this domain is therefore not visual realism but the geometric and photometric constraints imposed

by endoscopy: a wide depth range (0 to 30 mm during a complete procedure), weak texture, and strong specular reflections. The off-the-shelf evaluation reported here is deliberately a baseline; the natural next step is to use DRENDS to develop and assess surgical-specific models that combine the relative backbones already capable of recovering geometry with the scale-recovery strategies this benchmark identifies as the missing ingredient.

DARES, EndoDAC, Endo-FASt3r and EndoUFM adapt, through low-rank schemes and self-supervised photometric training on endoscopic video fine-tune the same backbones evaluated above, yet none of them improves on the backbone it derives from: once aligned they span 7.35 to 7.56 mm MAE on ex vivo and 12.22 to 13.14 mm on Phantom, against 6.75 and 10.48 mm for DAMv2 and 4.8 and 7.2 mm for the stereo methods. Their performance is bounded by the domain in which the adaptation was learnt: these models were adapted and validated on endoscopic sequences whose depth range, calibration and acquisition differ from those of DRENDS, so the adapted representation is evaluated here outside the distribution that produced it and the advantage it holds within that distribution does not follow it across. On top of this, a photometric objective constrains geometry only up to an unknown scale, so adaptation of this kind cannot supply the metric anchor the monocular family lacks, which is why the raw predictions remain in arbitrary inverse-depth units.

5.5 Discussion and Conclusion

This work introduced DRENDS, a high-resolution stereo-endoscopic dataset with per-frame metric depth derived from a synchronised time-of-flight sensor, explicitly capturing tissue manipulation and non-rigid deformations. By providing calibrated multi-modal recordings, precise timestamps, and an open processing pipeline for rectification, occlusion handling, and depth-map generation, DRENDS addresses a key limitation of existing resources: the scarcity of depth ground truth in dynamic surgical scenarios. At the same time, the dataset reflects a set of deliberate design choices. Depth annotations inherit the physical constraints of active time-of-flight sensing, including increased noise at depth discontinuities, sensitivity to surface reflectance, and partial occlusions during tool–tissue interaction. The measurements are bounded by the manufacturer-specified accuracy of the sensor and are further stabilised during preprocessing through temporal and spatial filtering. Rather than attempting to compensate for these effects using an explicit parametric noise model, the processing pipeline applies lightweight filtering designed to suppress measurement noise while preserving the natural characteristics of the sensor observations.

Furthermore, the data were acquired from a single experimental setup involving phantom and ex vivo tissue, which limits the diversity of anatomical structures across specimens. Moreover, DRENDS does not aim to exhaustively represent the full diversity of surgical procedures or anatomies; instead, it focuses on controlled phantom and ex vivo settings that emphasise clinically relevant depth variation, specular surfaces, and deformable tissue. While these scenarios do not fully capture in vivo physiological effects, they enable systematic and reproducible analysis of metric accuracy and temporal consistency under conditions that are difficult to isolate in unconstrained clinical data, positioning DRENDS as a principled step toward bridging the gap between synthetic benchmarks and in vivo deployment.

We anticipate that DRENDS will accelerate research in (i) metric depth estimation with reliable scale, (ii) temporally coherent depth prediction under deformation, and (iii) hybrid approaches that combine learning with surgical priors and multi-sensor constraints. By making the dataset and processing tools publicly available, we aim to support reproducible comparisons and encourage the development of methods that can translate into dependable 3D perception for robotic-assisted surgery.

Chapter 6

Conclusions

This work positions surgical data science as a cornerstone of robotic minimally invasive procedures (MIP), in which clinical decision-making and autonomous or semi-autonomous assistance depend on the intra-operative interpretation of complex surgical data under severe constraints. Within this setting, visual surgical perception is central to extracting an interpretation coherent with patient anatomy and the surgical workflow. Unlike conventional computer vision settings, endoscopic imaging operates in environments characterised by limited and viewpoint-dependent observations, non-rigid tissue deformation, frequent occlusions, specular reflections, and a lack of reliable metric ground truth. Within this context, visual perception cannot be reduced to isolated tasks such as object detection, tracking, or depth estimation; rather, it requires the construction of spatially and temporally consistent representations that support geometric reasoning across time and viewpoints.

This work aimed to acquire a comprehensive understanding of dynamic surgical scenes to support modern MIP systems by improving visual surgical perceptual capabilities through artificial intelligence. Here, visual surgical perception has been defined as a structured, multi-level problem, in which different representations enable different forms of reasoning and are inherently interdependent. Representations in this context were defined by the information extracted from different computer vision tasks that describe the surgical scene. Existing approaches often address these tasks in isolation, leading to fragile pipelines that struggle to generalise under deformation, scale ambiguity, and data scarcity. In parallel, progress in the field is further hindered by the limited availability of realistic datasets and by evaluation protocols that insufficiently capture metric accuracy and temporal consistency. These challenges identified key gaps in the field that this thesis sought to address.

6.1 Contributions

The contributions of this thesis arose from the research questions (RQs) introduced in Chapter 1, which progressively narrowed the problem of visual surgical perception from conceptual foundations to algorithmic and representational solutions. The analysis of the literature presented in the same chapter demonstrated that modern MIP systems strongly rely on visual perception and the ability to extract structured representations that support geometric reasoning and temporal integration. Artificial intelligence was therefore framed not as a replacement for clinical expertise, but as the studied mechanism for constructing and maintaining coherent scene representations under intra-operative constraints. These form the core reasoning and motivation behind each contribution in this thesis.

- *A representation-centric view of visual surgical perception, framing detection, tracking, depth, and reconstruction as coupled problems rather than isolated tasks.* **RQa** asked how computer vision tasks, through the representations they produce, define the limitations of visual surgical perception. Chapter 2 formalised visual surgical perception as a multi-level representation problem and proposed a structured framework linking computer vision tasks to abstraction levels in surgical data. This framework clarified how local features, object-level abstractions, correspondences, motion models, and reconstruction interact; it also exposed critical gaps in how existing approaches couple these levels and raise subsequent research questions. The central finding is that meaningful spatial understanding emerges from the dependencies between representations, rather than isolated task performance, and that the limitations of the field are therefore stated most usefully at the couplings between levels, since the output of one level is the input the next requires. This is a conceptual contribution corroborated by the three chapters that follow, each of which addresses a coupling the framework identifies, rather than one established by an experiment of its own.

- *Robust object representations for surgical scenes through a transformer-based detector designed to address scale variability and visual ambiguity via a stratified embedding space.* **RQb** asked how instruments and tissues can be represented so that their identity is recovered reliably under scale variation and visual similarity, at the rate an intraoperative system requires. Chapter 3 demonstrated that object representations can be achieved through transformer-based modelling with stratified embeddings designed to handle scale variability and

visual ambiguity. The detector reaches 0.945 $mAP_{[50]}$ at 32 FPS on m2cai16-tool-locations, ahead of every state-of-the-art and anchor-free method compared (Table 3.1), and improves average precision in every tool category over the baseline detector, from 2.7% for the irrigator to 11% for the hook (Table 3.2). The embedding space separates by class only when the contrastive term is included (Figure 3.6), which is what carries the claim that identity is recovered through the representation rather than through the classifier alone.

- *Render-based implicit scene representations that unify pixel-level appearance, 2D correspondences, 3D motion, and reconstruction in dynamic, deformable environments.* **RQc** asked how correspondences on deforming tissue can be maintained across a whole sequence rather than between neighbouring frames, and how camera motion can be separated from tissue motion. Chapter 4 introduced render-based implicit representations capable of jointly modelling pixel-level appearance, 2D correspondences, 3D motion, and scene structure. By embedding geometric constraints directly into the optimisation process, the proposed approach enabled temporally coherent tracking of deforming tissue and provided a principled mechanism for separating camera motion from scene deformation. It reduces the 2D tracking error to 13.925 px against 26.863 px for the closest test-time optimisation method and 31.793 px for OmniMotion, and tracks in 3D to 6.202 mm, which neither of those methods produces at all (Table 4.1).

- *A principled dataset and evaluation framework enabling metric and temporally consistent depth analysis under realistic tissue deformation.* **RQd** asked what is required to measure metric depth accuracy and temporal consistency in deformable surgical scenes, and how current models perform when measured that way. Chapter 5 presents the DRENDS dataset and its associated evaluation protocol, addressing a critical bottleneck in benchmarking surgical perception methods. Measured that way, no monocular model recovers metric depth out of the box, with errors of hundreds of millimetres before alignment, while stereo methods reach approximately 18 mm on the phantom subset and 12 mm ex vivo; after per-sequence alignment every model converges to low-millimetre error, showing that the gap is one of scale rather than of structure (Tables 5.2 and 5.3).

Overall, this thesis demonstrates that many limitations in surgical perception do not stem from individual algorithms, but from how visual information is represented and coupled across tasks. It further shows that different levels of abstraction

in scene representation enable distinct reasoning capabilities, and that meaningful spatial and temporal understanding emerges from the dependencies between these levels. It contributes with specific methods that efficiently create and integrate representations, surpassing state-of-the-art methods and with the potential to enable complex downstream reasoning. Finally, it also addressed the lack of assets that push the boundaries of the research by providing a unique dataset.

6.2 Limitations

While this thesis advances visual endoscopic perception through improved representations, integration strategies, and evaluation under realistic constraints, several limitations remain. These limitations are not confined to individual components of the perceptual pipeline but arise from shared assumptions, design choices, and the inherent complexity of surgical environments. A critical discussion of these aspects is essential to contextualise the scope of the contributions and to identify directions for future research.

6.2.1 Dependence on Frame-Level Supervision and Limited Temporal Context

A recurring limitation across the presented methods is the reliance on predominantly frame-wise or short-horizon supervision. Although temporal consistency is implicitly leveraged through tracking and test-time optimisation, explicit long-term temporal modelling remains limited. In particular, the object detection framework operates on individual frames, and misclassifications often arise in visually ambiguous situations where temporal context could provide disambiguating cues. Similarly, while the render-based tracking framework encodes motion through implicit representations, its performance ultimately depends on the quality of short-term correspondences. These observations highlight the need for stronger temporal priors and supervision strategies that extend beyond local frame-to-frame consistency.

6.2.2 Assumptions in Canonical and Implicit Representations

The render-based implicit scene representations introduced in this thesis rely on the assumption that deformable scenes can be mapped to a consistent canonical space within bounded regions. While this assumption enables stable tracking and

reconstruction in many scenarios, it may limit robustness under extreme deformations, prolonged occlusions, or topological changes. In such cases, distortions in the canonical mapping can affect convergence and tracking accuracy. More expressive representations that better preserve geodesic structure or adaptively update canonical topology could mitigate these limitations, but were beyond the scope of this work.

6.2.3 Dataset Scope and Generalisation

Although DRENDS captures realistic tissue deformation induced by instrument manipulation, it does not aim to represent the full diversity of surgical anatomies, procedures, or in vivo physiological effects. The dataset focuses on controlled phantom and ex vivo scenarios to enable systematic analysis of metric accuracy and temporal consistency. As a result, conclusions drawn from benchmarking on DRENDS may not directly translate to all clinical settings. Nevertheless, this trade-off was a deliberate design choice to prioritise reproducibility and controlled evaluation over breadth, and it reflects a broader challenge in surgical data science.

6.2.4 Computational Considerations

Several of the proposed methods, particularly those based on test-time optimisation and rendering, involve non-trivial computational costs. While these approaches demonstrate strong performance and flexibility, their current implementations may limit real-time deployment in resource-constrained clinical environments. Further work is required to explore hybrid strategies that retain the representational advantages of implicit and rendering-based methods while improving computational efficiency.

6.3 Future Research Directions

The limitations identified in this thesis highlight several promising directions for future research, many of which extend naturally from the representation-centric perspective adopted throughout this work. Rather than addressing individual tasks in isolation, these directions emphasise the development of perceptual systems that more tightly integrate spatial, temporal, and metric reasoning under realistic surgical constraints.

6.3.1 Semantic–Temporal Integration Across Perceptual Representations

A central opportunity for future work lies in incorporating semantic object information directly into render-based scene representations, enabling joint reasoning across temporal horizons about appearance, geometry, and identity. For example, object-level embeddings or class-conditioned features could be incorporated into implicit neural representations to guide correspondence estimation, disambiguate visually similar structures, and maintain semantic consistency under occlusion or deformation. Such integration would allow semantic priors to influence geometric optimisation, potentially reducing drift, improving long-term stability, and enabling more coherent interaction between detection, tracking, and reconstruction.

6.3.2 Adaptive and Topology-Aware Scene Representations

The render-based implicit representations introduced in this thesis rely on bounded canonical mappings that assume moderate deformation and stable topology. Future work could investigate more adaptive representations capable of handling extreme deformations, topological changes, or prolonged occlusions. Possible directions include topology-aware canonical spaces, dynamic reparameterisation strategies, or hybrid implicit–explicit representations that combine neural fields with surface-based or graph-based models. These extensions could improve convergence properties and preserve geodesic relationships under complex tissue motion, thereby strengthening the link between tracking, reconstruction, and metric reasoning.

6.3.3 Efficient and Deployable Visual Perceptual Systems

Finally, bridging the gap between expressive representations and deployable systems remains an open challenge. While rendering-based and test-time optimisation methods offer powerful modelling capabilities, their computational demands may limit clinical applicability. Future research should explore model compression, amortised optimisation, and hybrid inference strategies that retain the benefits of implicit representations while enabling near real-time performance. Addressing these challenges is critical for transitioning from research prototypes to practical perception modules in robotic-assisted surgery.

6.4 Final Remarks

This thesis has shown that the understanding of surgical scenes in computer-assisted and robotic minimally invasive interventions is fundamentally shaped by how visual information is abstracted, represented, and integrated across levels of perception. While strong relationships exist among data abstraction, computer vision tasks, and the requirements for automation in RMIS and CAI, these relationships are rarely made explicit in existing frameworks. By adopting a representation-centric perspective, this work establishes a coherent and quantitative link between surgical data, the abstraction levels at which it is processed, and the perceptual capabilities required to support spatial and temporal reasoning under realistic constraints. In doing so, the thesis not only advances individual components of the perceptual pipeline but also provides a structured view of how object-level semantics, geometric representations, and dynamic scene modelling can be combined to address the demands of advanced computer and robotic-assisted procedures. Taken together, the four questions are answered as follows: the limitations of the field are stated most usefully at the couplings between representation levels rather than within tasks; identity can be recovered through a supervised object embedding at real-time rates; correspondences can be held across a whole sequence, and camera motion separated from tissue motion, by an invertible representation optimised at test time; and metric depth in deformable scenes can be measured once ground truth and a protocol exist, at which point the gap between current models and clinical use proves to be one of scale rather than of structure. Each answer is bounded by the limitations set out above. It is hoped that this perspective will inform future research toward more integrated, interpretable, and dependable perception systems for robotic-assisted surgery.

Bibliography

[1] World Health Organization, *Who guidelines for safe surgery: Safe surgery saves lives*, Guideline, World Health Organization, Geneva, Switzerland, 2009, ISBN: 9789241598552. [Online]. Available: https://www.who.int/publications/i/item/9789241598552.

[2] NHS Digital, *National clinical coding standards: Opcs-4.10*, Version 4.10 of the OPCS Classification of Interventions and Procedures, NHS Digital, Leeds, United Kingdom, 2024. [Online]. Available: https://classbrowser.nhs.uk/ref_books/OPCS-4.10_NCCS-2024.pdf.

[3] J. G. Meara, A. J. M. Leather, L. Hagander, B. C. Alkire, N. Alonso, E. A. Ameh, S. W. Bickler, L. Conteh, A. J. Dare, J. Davies, E. D. Mérisier, S. El-Halabi, P. E. Farmer, A. Gawande, R. Gillies, S. L. M. Greenberg, C. E. Grimes, R. L. Gruen, E. A. Ismail, T. B. Kamara, C. Lavy, G. Lundeg, N. C. Mkandawire, N. P. Raykar, J. N. Riesel, E. Rodas, J. Rose, N. Roy, M. G. Shrime, R. Sullivan, S. Verguet, D. Watters, T. G. Weiser, I. H. Wilson, G. Yamey, and W. Yip, "Global surgery 2030: Evidence and solutions for achieving health, welfare, and economic development," *The lancet*, vol. 386, no. 9993, pp. 569–624, 2015.

[4] M. Tonutti, D. S. Elson, G.-Z. Yang, A. W. Darzi, and M. H. Sodergren, "The role of technology in minimally invasive surgery: State of the art, recent developments and future directions," *Postgraduate medical journal*, vol. 93, no. 1097, pp. 159–167, 2017.

[5] C. D'Ettorre, A. Mariani, A. Stilli, F. Rodriguez y Baena, P. Valdastri, A. Deguet, P. Kazanzides, R. H. Taylor, G. S. Fischer, S. P. DiMaio, A. Menciassi, and D. Stoyanov, "Accelerating surgical robotics research: A review of 10 years with the da vinci research kit," *IEEE Robotics & Automation Magazine*, vol. 28, no. 4, pp. 56–78, 2021.

[6] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic,

H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G. D. Hager, and P. Jannin, "Surgical data science for next-generation interventions," *Nature Biomedical Engineering*, vol. 1, no. 9, pp. 691–696, 2017.

[7] T. Robinson and G. Stiegmann, "Minimally invasive surgery," *Endoscopy*, vol. 36, no. 01, pp. 48–51, 2004.

[8] M. J. Mack, "Minimally invasive and robotic surgery," *Jama*, vol. 285, no. 5, pp. 568–572, 2001.

[9] L. Wyld, R. A. Audisio, and G. J. Poston, "The evolution of cancer surgery and future perspectives," *Nature reviews Clinical oncology*, vol. 12, no. 2, pp. 115–124, 2015.

[10] A. Obermair, R. Asher, R. Pareja, M. Frumovitz, A. Lopez, R. Moretti-Marques, G. Rendon, R. Ribeiro, A. Tsunoda, V. Behan, A. Buda, M. Q. Bernadini, H. Zhao, M. Vieira, J. Walker, N. M. Spirtos, S. Yao, N. Chetty, T. Zhu, D. Isla, M. Tamura, J. Nicklin, K. P. Robledo, V. Gebski, R. L. Coleman, G. Salvo, and P. T. Ramirez, "Incidence of adverse events in minimally invasive vs open radical hysterectomy in early cervical cancer: Results of a randomized controlled trial," *American journal of obstetrics and gynecology*, vol. 222, no. 3, 249–e1, 2020.

[11] P. C. De Groen, "History of the endoscope [scanning our past]," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1987–1995, 2017.

[12] R. Hargest, "Five thousand years of minimal access surgery: 3000bc to 1850: Early instruments for viewing body cavities," *Journal of the Royal Society of Medicine*, vol. 113, no. 12, pp. 491–496, 2020.

[13] R. Hargest, "Five thousand years of minimal access surgery: 1850 to 1990: Technological developments," *Journal of the Royal Society of Medicine*, vol. 114, no. 1, pp. 19–29, 2021.

[14] R. Hargest, "Five thousand years of minimal access surgery: 1990–present: Organisational issues and the rise of the robots," *Journal of the Royal Society of Medicine*, vol. 114, no. 2, pp. 69–76, 2021.

[15] S. Cousins, N. S. Blencowe, and J. M. Blazeby, "What is an invasive procedure? a definition to inform study design, evidence synthesis and research tracking," *BMJ open*, vol. 9, no. 7, e028576, 2019.

[16] E. Ray-Offor and R. J. Rosenthal, *Principles and practice of laparoscopic surgery*, 2025.

[17] J. G. Hunter, D. H. Spight, C. Sandone, and J. E. Fairman, "Principles of room setup and port placement for minimally invasive surgery," in *Atlas of Minimally Invasive Surgical Operations*. New York, NY: McGraw-Hill Education, 2018. [Online]. Available: accesssurgery.mhmedical.com/content.aspx?aid=1162529404.

[18] A. Haycock, J. Cohen, B. P. Saunders, P. B. Cotton, and C. B. Williams, *Cotton and Williams' practical gastrointestinal endoscopy: the fundamentals*. John Wiley & Sons, 2013.

[19] Y. Hori, "Diagnostic laparoscopy guidelines," *Surgical Endoscopy*, vol. 22, no. 5, pp. 1353–1383, 2008, Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) Guidelines Committee. DOI: 10.1007/s00464-008-9759-5.

[20] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario, "Medical robotics and computer-integrated surgery," in *Springer Handbook of Robotics*, B. Siciliano and O. Khatib, Eds. Cham: Springer International Publishing, 2016, pp. 1657–1684, ISBN: 978-3-319-32552-1. DOI: 10.1007/978-3-319-32552-1_63. [Online]. Available: https://doi.org/10.1007/978-3-319-32552-1_63.

[21] J. H. Siewerdsen, "Image-guided surgery: Technology and opportunities for medical physics," *Medical physics*, vol. 52, no. 10, e70049, 2025.

[22] K. Cleary and T. M. Peters, "Image-guided interventions: Technology review and clinical applications," *Annual review of biomedical engineering*, vol. 12, no. 1, pp. 119–142, 2010.

[23] M. H. A. Nasab, *Handbook of robotic and image-guided surgery*. Elsevier, 2019.

[24] Moon Surgical, *Maestro system*, https://moonsurgical.com/, Accessed: 2026-08-27, 2026.

[25] NVIDIA Corporation, *Nvidia holoscan*, https://www.nvidia.com/en-gb/clara/holoscan/, Accessed: 2026-08-27, 2026.

[26] A. Kumar, N. Yadav, S. Singh, and N. Chauhan, "Minimally invasive (endoscopic-computer assisted) surgery: Technique and review," *Annals of maxillofacial surgery*, vol. 6, no. 2, pp. 159–164, 2016.

[27] S. Ali, "Where do we stand in ai for endoscopic image analysis? deciphering gaps and future directions," *npj Digital Medicine*, vol. 5, no. 1, p. 184, 2022.

[28] A. Handa, A. Gaidhane, and S. G. Choudhari, "Role of robotic-assisted surgery in public health: Its advantages and challenges," *Cureus*, vol. 16, no. 6, 2024.

[29] Intuitive Surgical, Inc., *Intuitive surgical*, https://www.intuitive.com/en-gb, Accessed: 2025-01-XX, 2025.

[30] Medtronic plc, *Hugo^{TM} robotic-assisted surgery system*, https://www.medtronic.com/covidien/en-gb/robotic-assisted-surgery/hugo-ras-system.html, Accessed: 2025-01-XX, 2025.

[31] CMR Surgical Ltd., *Cmr surgical*, https://cmrsurgical.com/, Accessed: 2025-01-XX, 2025.

[32] Asensus Surgical, Inc., *Senhance surgical system*, https://www.asensus.com/senhance, Accessed: 2025-01-XX, 2025.

[33] P. Kazanzides, Z. Chen, A. Deguet, G. S. Fischer, R. H. Taylor, and S. P. DiMaio, "An open-source research kit for the da vinci® surgical system," in *2014 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2014, pp. 6434–6439.

[34] D. Jayne, A. Pigazzi, H. Marshall, J. Croft, N. Corrigan, J. Copeland, P. Quirke, N. West, T. Rautio, N. Thomassen, H. Tilney, M. Gudgeon, P. P. Bianchi, R. Edlin, C. Hulme, and J. Brown, "Effect of robotic-assisted vs conventional laparoscopic surgery on risk of conversion to open laparotomy among patients undergoing resection for rectal cancer: The rolarr randomized clinical trial," *Jama*, vol. 318, no. 16, pp. 1569–1580, 2017.

[35] E. Birgin, M. Heibel, S. Hetjens, E. Rasbach, C. Reissfelder, P. Téoule, and N. N. Rahbari, "Robotic versus laparoscopic hepatectomy for liver malignancies (roc'n'roll): A single-centre, randomised, controlled, single-blinded clinical trial," *The Lancet Regional Health–Europe*, vol. 43, p. 100 972, 2024.

[36] J. Troccaz, G. Dagnino, and G.-Z. Yang, "Frontiers of medical robotics: From concept to systems to clinical translation," *Annual review of biomedical engineering*, vol. 21, no. 1, pp. 193–218, 2019.

[37] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastri, "Autonomy in surgical robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, no. 1, pp. 651–679, 2021.

[38] B. T. Ostrander, D. Massillon, L. Meller, Z.-Y. Chiu, M. Yip, and R. K. Orosco, "The current state of autonomous suturing: A systematic review," *Surgical Endoscopy*, vol. 38, no. 5, pp. 2383–2397, 2024.

[39] E. Tagliabue, D. Dall'Alba, M. Pfeiffer, M. Piccinelli, R. Marin, U. Castellani, S. Speidel, and P. Fiorini, "Data-driven intra-operative estimation of anatomical attachments for autonomous tissue dissection," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1856–1863, 2021.

[40] S. Eslamian, L. A. Reisner, and A. K. Pandya, "Development and evaluation of an autonomous camera control algorithm on the da vinci surgical system," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 16, no. 2, e2036, 2020.

[41] J. Hu, D. Jones, M. R. Dogar, and P. Valdastri, "Occlusion-robust autonomous robotic manipulation of human soft tissues with 3-d surface feedback," *IEEE Transactions on Robotics*, vol. 40, pp. 624–638, 2023.

[42] Y. Kassahun, B. Yu, A. T. Tibebu, D. Stoyanov, S. Giannarou, J. H. Metzen, and E. Vander Poorten, "Surgical robotics beyond enhanced dexterity instrumentation: A survey of machine learning techniques and their role in intelligent and autonomous surgical actions," *International journal of computer assisted radiology and surgery*, vol. 11, no. 4, pp. 553–568, 2016.

[43] R. H. Mithany, S. Aslam, S. Abdallah, M. Abdelmaseeh, F. Gerges, M. S. Mohamed, M. Manasseh, A. Wanees, M. H. Shahid, M. S. Khalil, and N. Daniel, "Advancements and challenges in the application of artificial intelligence in surgical arena: A literature review," *Cureus*, vol. 15, no. 10, 2023.

[44] S. Datta, Y. Li, M. M. Ruppert, Y. Ren, B. Shickel, T. Ozrazgat-Baslanti, P. Rashidi, and A. Bihorac, "Reinforcement learning in surgery," *Surgery*, vol. 170, no. 1, pp. 329–332, 2021.

[45] X. Luo, K. Mori, and T. M. Peters, "Advanced endoscopic navigation: Surgical big data, methodology, and applications," *Annual review of biomedical engineering*, vol. 20, no. 1, pp. 221–251, 2018.

[46] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA, United States: MIT Press, 1982.

[47] L. Maier-Hein, M. Eisenmann, D. Sarikaya, K. März, T. Collins, A. Malpani, J. Fallert, H. Feussner, S. Giannarou, P. Mascagni, H. Nakawala, A. Park, C. Pugh, D. Stoyanov, S. S. Vedula, K. Cleary, G. Fichtinger, G. Forestier, B. Gibaud, T. Grantcharov, M. Hashizume, D. Heckmann-Nötzel, H. G. Kenngott, R. Kikinis, L. Mündermann, N. Navab, S. Onogur, T. Roß, R. Sznitman, R. H. Taylor, M. D. Tizabi, M. Wagner, G. D. Hager, T. Neumuth, N. Padoy, J. Collins, I. Gockel, J. Goedeke, D. A. Hashimoto, L. Joyeux, K.

Lam, D. R. Leff, A. Madani, H. J. Marcus, O. Meireles, A. Seitel, D. Teber, F. Ückert, B. P. Müller-Stich, P. Jannin, and S. Speidel, "Surgical data science–from concepts toward clinical translation," *Medical image analysis*, vol. 76, p. 102 306, 2022.

[48] P. Mascagni, D. Alapatt, L. Sestini, M. S. Altieri, A. Madani, Y. Watanabe, A. Alseidi, J. A. Redan, S. Alfieri, G. Costamagna, I. Boškoski, N. Padoy, and D. A. Hashimoto, "Computer vision in surgery: From potential to clinical value," *npj Digital Medicine*, vol. 5, no. 1, p. 163, 2022.

[49] R. Szeliski, *Computer Vision: Algorithms and Applications*, 2nd ed. Cham, Switzerland: Springer, 2022.

[50] A. A. Gumbs, V. Grasso, N. Bourdel, R. Croner, G. Spolverato, I. Frigerio, A. Illanes, M. Abu Hilal, A. Park, and E. Elyan, "The advances in computer vision that are enabling more autonomous actions in surgery: A systematic review of the literature," *Sensors*, vol. 22, no. 13, p. 4918, 2022.

[51] A. Schmidt, O. Mohareri, S. DiMaio, M. C. Yip, and S. E. Salcudean, "Tracking and mapping in medical computer vision: A review," *Medical Image Analysis*, vol. 94, p. 103 131, 2024.

[52] A. Richter, T. Steinmann, J.-C. Rosenthal, and S. J. Rupitsch, "Advances in real-time 3d reconstruction for medical endoscopy," *Journal of imaging*, vol. 10, no. 5, p. 120, 2024.

[53] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[54] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[55] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[56] 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.

[57] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3456–3465.

[58] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint detection and description of local features," *arXiv preprint arXiv:1905.03561*, 2019.

[59] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2d2: Reliable and repeatable detector and descriptor," *Advances in neural information processing systems*, vol. 32, 2019.

[60] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging*, vol. 36, no. 1, pp. 86–97, 2016.

[61] Y. Wang, Q. Sun, Z. Liu, and L. Gu, "Visual detection and tracking algorithms for minimally invasive surgical instruments: A comprehensive review of the state-of-the-art," *Robotics and Autonomous Systems*, vol. 149, p. 103 945, 2022.

[62] M. Ali, R. M. G. Pena, G. O. Ruiz, and S. Ali, "A comprehensive survey on recent deep learning-based methods applied to surgical data," *arXiv preprint arXiv:2209.01435*, 2022.

[63] D. Bouget, M. Allan, D. Stoyanov, and P. Jannin, "Vision-based and marker-less surgical tool detection and tracking: A review of the literature," *Medical image analysis*, vol. 35, pp. 633–654, 2017.

[64] F. Chadebecq, L. B. Lovat, and D. Stoyanov, "Artificial intelligence and automation in endoscopy and surgery," *Nature Reviews Gastroenterology & Hepatology*, vol. 20, no. 3, pp. 171–182, 2023.

[65] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578.

[66] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[67] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, Springer, 2020, pp. 213–229.

[68] Y. Liu, Z. Zhao, F. Chang, and S. Hu, "An anchor-free convolutional neural network for real-time surgical tool detection in robot-assisted surgery," *IEEE Access*, vol. 8, pp. 78 193–78 201, 2020.

[69] R. Anteby, N. Horesh, S. Soffer, Y. Zager, Y. Barash, I. Amiel, D. Rosin, M. Gutman, and E. Klang, "Deep learning visual analysis in laparoscopic surgery: A systematic review and diagnostic test accuracy meta-analysis," *Surgical endoscopy*, vol. 35, no. 4, pp. 1521–1533, 2021.

[70] M. Carstens, F. M. Rinner, S. Bodenstedt, A. C. Jenke, J. Weitz, M. Distler, S. Speidel, and F. R. Kolbinger, "The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science," *Scientific Data*, vol. 10, no. 1, pp. 1–8, 2023.

[71] A. Murali, D. Alapatt, P. Mascagni, A. Vardazaryan, A. Garcia, N. Okamoto, G. Costamagna, D. Mutter, J. Marescaux, B. Dallemagne, and N. Padoy, "The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark," *arXiv preprint arXiv:2312.12429*, 2023.

[72] D. Sarikaya, J. J. Corso, and K. A. Guru, "Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection," *IEEE transactions on medical imaging*, vol. 36, no. 7, pp. 1542–1549, 2017.

[73] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, and L. Fei-Fei, "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*, IEEE, 2018, pp. 691–699.

[74] B. Zhang, S. Wang, L. Dong, and P. Chen, "Surgical tools detection based on modulated anchoring network in laparoscopic videos," *Ieee Access*, vol. 8, pp. 23 748–23 758, 2020.

[75] B. Choi, K. Jo, S. Choi, and J. Choi, "Surgical-tools detection based on convolutional neural network in laparoscopic robot-assisted surgery," in *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, Ieee, 2017, pp. 1756–1759.

[76] C. S. Pradeep and N. Sinha, "Multi-tasking dssd architecture for laparoscopic cholecystectomy surgical assistance systems," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2022, pp. 1–4.

[77] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: Deconvolutional single shot detector," *arXiv preprint arXiv:1701.06659*, 2017.

[78] Z. Zhao, S. Voros, Z. Chen, and X. Cheng, "Surgical tool tracking based on two cnns: From coarse to fine," *The Journal of Engineering*, vol. 2019, no. 14, pp. 467–472, 2019.

[79] A. Vardazaryan, D. Mutter, J. Marescaux, and N. Padoy, "Weakly-supervised learning for tool localization in laparoscopic videos," in *International Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis*, Springer, 2018, pp. 169–179.

[80] P. Shi, Z. Zhao, S. Hu, and F. Chang, "Real-time surgical tool detection in minimally invasive surgery based on attention-guided convolutional neural network," *IEEE Access*, vol. 8, pp. 228 853–228 862, 2020.

[81] Y. Yang, Z. Zhao, P. Shi, and S. Hu, "An efficient one-stage detector for real-time surgical tools detection in robot-assisted surgery," in *Annual conference on medical image understanding and analysis*, Springer, 2021, pp. 18–29.

[82] T. Abdulbaki Alshirbaji, N. A. Jalal, H. Arabian, A. Battistel, P. Docherty, H. ElMoaqet, T. Neumuth, and K. Moeller, *Cholec80-boxes: Bounding-box labels for surgical tools in five cholecystectomy videos*, Zenodo, Sep. 2024. DOI: 10.5281/zenodo.13170928. [Online]. Available: https://doi.org/10.5281/zenodo.13170928.

[83] S. Kondo, "Lapformer: Surgical tool detection in laparoscopic surgical video using transformer architecture," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 9, no. 3, pp. 302–307, 2021.

[84] A. C. Jenke, S. Bodenstedt, F. R. Kolbinger, M. Distler, J. Weitz, and S. Speidel, "One model to use them all: Training a segmentation model with complementary datasets," *International journal of computer assisted radiology and surgery*, vol. 19, no. 6, pp. 1233–1241, 2024.

[85] P. M. Scheikl, S. Laschewski, A. Kisilenko, T. Davitashvili, B. Müller, M. Capek, B. P. Müller-Stich, M. Wagner, and F. Mathis-Ullrich, "Deep learning for semantic segmentation of organs and tissues in laparoscopic surgery," in *Current directions in biomedical engineering*, De Gruyter, vol. 6, 2020, p. 20 200 016.

[86] A. Attanasio, C. Alberti, B. Scaglioni, N. Marahrens, A. F. Frangi, M. Leonetti, C. S. Biyani, E. De Momi, and P. Valdastri, "A comparative study of spatio-temporal u-nets for tissue segmentation in surgical robotics," *IEEE Transactions on Medical Robotics and Bionics*, vol. 3, no. 1, pp. 53–63, 2021.

[87] S. G. Park, J. Park, H. R. Choi, J. H. Lee, S. T. Cho, Y. G. Lee, H. Ahn, and S. Pak, "Deep learning model for real-time semantic segmentation during intraoperative robotic prostatectomy," *European Urology Open Science*, vol. 62, pp. 47–53, 2024.

[88] Z. Gao, Q. Wu, Y. CHEN, C. Chen, H. C. Yip, W. C. W. Chu, and Q. Dou, "Temporal memory enhancement for semantic segmentation in surgical video," in *Medical Imaging with Deep Learning*, 2026. [Online]. Available: https://openreview.net/forum?id=arGD0IznGt.

[89] Y. Zhou, H. Badgery, M. Read, J. Bailey, and C. E. Davey, "Surgicalsemiseg: A semi-supervised framework for laparoscopic image segmentation," in *Medical Imaging with Deep Learning*, 2025.

[90] R. Zhou, D. Wang, H. Zhang, Y. Zhu, L. Zhang, T. Chen, W. Liao, and Z. Ye, "Vision techniques for anatomical structures in laparoscopic surgery: A comprehensive review," *Frontiers in Surgery*, vol. 12, p. 1 557 153, 2025.

[91] A. Schmidt, M. A. Karaoglu, S. Sinha, M. Jang, H.-G. Ha, K. Jung, K. Gu, I. Ullah, H. Lee, J. Šerých, M. Neoral, J. Matas, R. Zhou, W. He, A. Wang, H. Ren, B. Silva, S. Queirós, E. Lima, J. L. Vilaça, S. Kikuchi, A. Kouno, H. Matsuzaki, T. Li, Y. Chen, L. Li, X. Ma, X. Li, M. S. Zeinoddin, X. Wang, Z. Tandogdu, G. Shaw, E. Mazomenos, D. Stoyanov, Y. Chen, Z. Wu, A. Ladikos, S. DiMaio, S. E. Salcudean, and O. Mohareri, "Point tracking in surgery–the 2024 surgical tattoos in infrared (stir) challenge," *arXiv preprint arXiv:2503.24306*, 2025.

[92] N. Haouchine, W. Kuang, S. Cotin, and M. Yip, "Vision-based force feedback estimation for robot-assisted surgery using instrument-constrained biomechanical three-dimensional maps," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2160–2165, 2018.

[93] A. Schmidt and S. E. Salcudean, "Real-time rotated convolutional descriptor for surgical environments," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 279–289.

[94] G. Potje, F. Cadar, A. Araujo, R. Martins, and E. R. Nascimento, "Enhancing deformable local features by jointly learning to detect and describe keypoints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1306–1315.

[95] M. A. Karaoglu, V. Markova, N. Navab, B. Busam, and A. Ladikos, "Ride: Self-supervised learning of rotation-equivariant keypoint detection and invariant description for endoscopy," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2024, pp. 10 764–10 771.

[96] Y. Chu, H. Li, X. Li, Y. Ding, X. Yang, D. Ai, X. Chen, Y. Wang, and J. Yang, "Endoscopic image feature matching via motion consensus and global bilateral regression," *Computer Methods and Programs in Biomedicine*, vol. 190, p. 105 370, 2020.

[97] G. Zhang, G. Feng, F. He, and Z. Jiang, "Robust feature matching for vslam in non-rigid scenes," in *International Conference on Electronic Information Engineering and Computer Science (EIECS 2022)*, SPIE, vol. 12602, 2023, pp. 443–450.

[98] A. Schmidt, O. Mohareri, S. DiMaio, and S. E. Salcudean, "Fast graph refinement and implicit neural representation for tissue tracking," in *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, 2022, pp. 1281–1288.

[99] A. Lukezic, T. Vojir, L. ˇCehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6309–6318.

[100] Z. Liu, W. Gao, J. Zhu, Z. Yu, and Y. Fu, "Surface deformation tracking in monocular laparoscopic video," *Medical image analysis*, vol. 86, p. 102 775, 2023.

[101] V. Penza, X. Du, D. Stoyanov, A. Forgione, L. S. Mattos, and E. De Momi, "Long term safety area tracking (lt-sat) with online failure detection and recovery for robotic minimally invasive surgery," *Medical image analysis*, vol. 45, pp. 13–23, 2018.

[102] S. Ihler, F. Kuhnke, M.-H. Laves, and T. Ortmaier, "Self-supervised domain adaptation for patient-specific, real-time tissue tracking," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2020, pp. 54–64.

[103] J. Cartucho, A. Weld, S. Tukra, H. Xu, H. Matsuzaki, T. Ishikawa, M. Kwon, Y. E. Jang, K.-J. Kim, G. Lee, B. Bai, L. A. Kahrs, L. Boecking, S. Allmendinger, L. Müller, Y. Zhang, Y. Jin, S. Bano, F. Vasconcelos, W. Reiter, J. Hajek, B. Silva, E. Lima, J. L. Vilaça, S. Queirós, and S. Giannarou, "Surgt challenge: Benchmark of soft-tissue trackers for robotic surgery," *Medical image analysis*, vol. 91, p. 102 985, 2024.

[104] D. Recasens, J. Lamarca, J. M. Fácil, J. M. Montiel, and J. Civera, "Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7225–7232, 2021.

[105] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissanayake, "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.

[106] J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Tracking monocular camera pose and deformation for slam inside the human body," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2022, pp. 5278–5285.

[107] H. Zhou and J. Jayender, "Emdq-slam: Real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 331–340.

[108] R. Ma, R. Wang, Y. Zhang, S. Pizer, S. K. McGill, J. Rosenman, and J.-M. Frahm, "Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy," *Medical image analysis*, vol. 72, p. 102 100, 2021.

[109] A. Schmidt, O. Mohareri, S. DiMaio, and S. E. Salcudean, "Sendd: Sparse efficient neural depth and deformation for tissue tracking," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, pp. 238–248.

[110] A. Rau, B. Bhattarai, L. Agapito, and D. Stoyanov, "Bimodal camera pose prediction for endoscopy," *IEEE Transactions on Medical Robotics and Bionics*, vol. 5, no. 4, pp. 978–989, 2023.

[111] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[112] S. Leonard, A. Sinha, A. Reiter, M. Ishii, G. L. Gallia, R. H. Taylor, and G. D. Hager, "Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data," *IEEE transactions on medical imaging*, vol. 37, no. 10, pp. 2185–2195, 2018.

[113] O. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. Montiel, "Visual slam for handheld monocular endoscope," *IEEE transactions on medical imaging*, vol. 33, no. 1, pp. 135–146, 2013.

[114] A. Rau, S. Bano, Y. Jin, P. Azagra, J. Morlana, R. Kader, E. Sanderson, B. J. Matuszewski, J. Y. Lee, D.-J. Lee, E. Posner, N. Frank, V. Elangovan, S. Raviteja, Z. Li, J. Liu, S. Lalithkumar, M. Islam, H. Ren, L. B. Lovat, J. M. Montiel, and D. Stoyanov, "Simcol3d—3d reconstruction during colonoscopy challenge," *Medical Image Analysis*, vol. 96, p. 103 195, 2024.

[115] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[116] K. B. Ozyoruk, G. I. Gokceler, T. L. Bobrow, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira, H. Sahin, H. Araujo, H. Alexandrino, N. J. Durr, H. B. Gilbert, and M. Turan, "Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," *Medical Image Analysis*, vol. 71, p. 102 058, 2021, ISSN: 1361-8415.

[117] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[118] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[119] Y. Wang, Y. Long, S. H. Fan, and Q. Dou, "Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery," in *International conference on medical image computing and computer-assisted intervention*, Springer, 2022, pp. 431–441.

[120] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6197–6206.

[121] M. Hayoz, C. Hahne, T. Kurmann, M. Allan, G. Beldi, D. Candinas, P. Márquez-Neila, and R. Sznitman, "Online 3d reconstruction and dense tracking in endoscopic videos," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2024, pp. 444–454.

[122] Q. Wang, Y.-Y. Chang, R. Cai, Z. Li, B. Hariharan, A. Holynski, and N. Snavely, "Tracking everything everywhere all at once," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 795–19 806.

[123] B. Zhan, W. Zhao, Y. Fang, B. Du, F. Vasconcelos, D. Stoyanov, D. S. Elson, and B. Huang, "Tracking everything in robotic-assisted surgery," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2025, pp. 1–7.

[124] M. Kam, S. Wei, J. D. Opfermann, H. Saeidi, M. H. Hsieh, K. C. Wang, J. U. Kang, and A. Krieger, "Autonomous system for vaginal cuff closure via model-based planning and markerless tracking techniques," *IEEE Robotics and Automation Letters*, vol. 8, no. 7, pp. 3916–3923, 2023.

[125] J. Ge, H. Saeidi, J. D. Opfermann, A. S. Joshi, and A. Krieger, "Landmark-guided deformable image registration for supervised autonomous robotic tumor resection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2019, pp. 320–328.

[126] R. Y. Tsai and R. K. Lenz, "A new technique for fully autonomous and efficient 3d robotics hand/eye calibration," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 345–358, 1989. DOI: 10.1109/70.34770.

[127] G. Loza, M. MAGRO, and B. Calmé, *Drends: A dataset for depth in robotic endoscopy with dynamic scenarios*, version 1.0, Zenodo, Nov. 2025. DOI: 10.5281/zenodo.17598453. [Online]. Available: https://doi.org/10.5281/zenodo.17598453.

[128] L. Maier-Hein, P. Mountney, A. Bartoli, H. Elhawary, D. Elson, A. Groch, A. Kolb, M. Rodrigues, J. Sorger, S. Speidel, and D. Stoyanov, "Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery," *Medical image analysis*, vol. 17, no. 8, pp. 974–996, 2013.

[129] E. Edwards, D. Psychogyios, L. Maier-Hein, S. Speidel, and D. Stoyanov, "Serv-ct: A disparity dataset from cone-beam ct for validation of endoscopic 3d reconstruction," *Medical image analysis*, vol. 76, p. 102 302, 2024. DOI: 10.5522/04/26352199.v1.

[130] V. Penza, A. S. Ciullo, S. Moccia, L. S. Mattos, and E. De Momi, "Endoabs dataset: Endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 14, no. 5, e1926, 2018.

[131] M. Allan, J. Mcleod, C. Wang, J. C. Rosenthal, Z. Hu, N. Gard, P. Eisert, K. X. Fu, T. Zeffiro, W. Xia, Z. Zhu, H. Luo, F. Jia, X. Zhang, X. Li, L. Sharan, T. Kurmann, S. Schmid, R. Sznitman, D. Psychogyios, M. Azizian, D. Stoyanov, L. Maier-Hein, and S. Speidel, "Stereo correspondence and reconstruction of endoscopic data challenge," *arXiv preprint arXiv:2101.01133*, 2021. DOI: 10.48550/arXiv.2101.01133.

[132] T. L. Bobrow, M. Golhar, R. Vijayan, V. S. Akshintala, J. R. Garcia, and N. J. Durr, "Colonoscopy 3d video dataset with paired depth from 2d-3d registration," *Medical Image Analysis*, vol. 90, p. 102 956, 2023.

[133] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[134] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PmLR, 2020, pp. 1597–1607.

[135] M. Sahu, A. Mukhopadhyay, A. Szengel, and S. Zachow, "Addressing multi-label imbalance problem of surgical tool detection using cnn," *International journal of computer assisted radiology and surgery*, vol. 12, no. 6, pp. 1013–1020, 2017.

[136] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[137] M. A. Teevno, G. Ochoa-Ruiz, and S. Ali, "A semi-supervised teacher-student framework for surgical tool detection and localization," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 11, no. 4, pp. 1033–1041, 2023.

[138] Y. Song, J. Lei, Z. Wang, L. Liu, and K. Daniilidis, "Track everything everywhere fast and robustly," in *European Conference on Computer Vision*, Springer, 2024, pp. 343–359.

[139] R. Li, H. Gao, M. Tancik, and A. Kanazawa, "Nerfacc: Efficient sampling accelerates nerfs," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 18 537–18 546.

[140] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*, Springer, 2020, pp. 402–419.

[141] N. Karaev, Y. Makarov, J. Wang, N. Neverova, A. Vedaldi, and C. Rupprecht, "Cotracker3: Simpler and better point tracking by pseudo-labelling real videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 6013–6022.

[142] M. Neoral, J. Šerỳch, and J. Matas, "Mft: Long-term tracking of every pixel," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6837–6847.

[143] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5470–5479.

[144] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[145] A. Cao and J. Johnson, "Hexplane: A fast representation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 130–141.

[146] A. Schmidt, O. Mohareri, S. P. DiMaio, and S. E. Salcudean, "Surgical tattoos in infrared: A dataset for quantifying tissue tracking and mapping," *IEEE Transactions on Medical Imaging*, vol. 43, no. 7, pp. 2634–2645, 2024.

[147] J. Cartucho, A. Weld, S. Tukra, H. Xu, H. Matsuzaki, T. Ishikawa, M. Kwon, Y. E. Jang, K.-J. Kim, G. Lee, B. Bai, L. A. Kahrs, L. Boecking, S. Allmendinger, L. Müller, Y. Zhang, Y. Jin, S. Bano, F. Vasconcelos, W. Reiter, J. Hajek, B. Silva, E. Lima, J. L. Vilaça, S. Queirós, and S. Giannarou, "Surgt challenge: Benchmark of soft-tissue trackers for robotic surgery," *Medical Image Analysis*, vol. 91, p. 102 985, 2024. DOI: 10.1016/j.media.2023.102985.

[148] K.-H. Oh, L. Borgioli, A. Mangano, V. Valle, M. Di Pangrazio, F. Toti, G. Pozza, L. Ambrosini, A. Ducas, M. Žefran, L. Chen, and P. C. Giulianotti, "Expanded comprehensive robotic cholecystectomy dataset," *Journal of Medical Robotics Research*, vol. 10, no. 03n04, p. 2 550 006, 2025. DOI: 10.1142/S2424905X25500060.

[149] R. Zhou, W. He, A. Wang, Q. Yao, H. Hu, J. Wang, X. Zhang, and H. Ren, "Endo-ttap: Robust endoscopic tissue tracking via multi-facet guided attention and hybrid flow-point supervision," *arXiv preprint arXiv:2503.22394*, 2025.

[150] W. Burger, "Zhang's camera calibration algorithm: In-depth tutorial and implementation," *HGB16-05*, pp. 1–6, 2016.

[151] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, IEEE, 1998, pp. 839–846.

[152] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024. [Online]. Available: https://arxiv.org/abs/2408.00714.

[153] G. He, T. Lu, H. Li, J. Lu, and H. Zhu, "Patch tensor decomposition and non-local means filter-based hybrid asl image denoising," *Journal of Neuroscience Methods*, vol. 370, p. 109 488, 2022.

[154] C. Budd and T. Vercauteren, "Transferring relative monocular depth to surgical vision with temporal consistency," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2024, pp. 692–702.

[155] A. Bochkovskiy, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. Richter, and V. Koltun, "Depth pro: Sharp monocular metric depth in less than a second," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=aueXfY0Clv.

[156] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[157] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21 875–21 911, 2024.

[158] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," 2023. DOI: 10.48550/ARXIV.2302.12288.

[159] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*, IEEE, 2021, pp. 218–227.

[160] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying flow, stereo and depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[161] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 5249–5260.

[162] M. Sheikh Zeinoddin, C. Lena, J. Qu, L. Carlini, M. Magro, S. Kim, E. De Momi, S. Bano, M. Grech-Sollars, E. Mazomenos, *et al.*, "Dares: Depth anything in robotic endoscopic surgery with self-supervised vector-lora of the foundation model," in *European Conference on Computer Vision*, Springer, 2024, pp. 1–11.

[163] B. Cui, M. Islam, L. Bai, A. Wang, and H. Ren, "Endodac: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2024, pp. 208–218.

[164] M. Sheikh Zeinoddin, M. I. Hoque, Z. Tandogdu, G. L. Shaw, M. J. Clarkson, E. B. Mazomenos, and D. Stoyanov, "Endo-fast3r: Endoscopic foundation model adaptation for structure from motion," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2025, pp. 117–126.

[165] X. Yao, B. Liu, B. Li, J. Wang, J. Yue, and F. Zhou, "Endoufm: Utilizing foundation models for monocular depth estimation of endoscopic images," *Neural Networks*, p. 109 141, 2026.

[166] B. Cui, M. Islam, L. Bai, and H. Ren, "Surgical-dino: Adapter learning of foundation models for depth estimation in endoscopic surgery," *International Journal of Computer Assisted Radiology and Surgery*, vol. 19, no. 6, pp. 1013–1020, 2024.

[167] X. Wang, J. Xu, S. Zhang, B. Huang, D. Stoyanov, and E. B. Mazomenos, "Stereomamba: Real-time and robust intraoperative stereo disparity estimation via long-range spatial dependencies," *IEEE Robotics and Automation Letters*, vol. 10, no. 10, 2025.

Appendix A

Complementary material Ch. 3

Table A.1: **Grid search for important hyper-parameters and ablation study.**

Model	**mAP**$_{[50]}$	**mAP**$_{[75]}$	**FPS**	**Queries**
DETR	0.901	0.552	36	32
DETR	0.896	0.507	36	64
DETR	0.857	0.483	36	100
Model	**mAP**$_{[50]}$	**mAP**$_{[75]}$	**FPS**	**Layers Tx-e,Tx-d**
DETR	0.895	0.550	36	[6,6]
DETR	0.837	0.470	38	[5,5]
DETR	0.856	0.450	41	[4,4]
DETR	0.774	0.401	45	[3,3]
DETR	0.715	0.220	49	[2,2]
DETR	0.624	0.153	51	[1,1]
Model	**mAP**$_{[50]}$	**mAP**$_{[75]}$	**FPS**	**scales,feats**
DETR + MS	0.908	0.488	32	[4,26]
DETR + MS	0.905	0.488	26	[6,26]
DETR + MS	0.924	0.521	24	[8,26]
DETR + MS	0.899	0.443	21	[8,14]
Model	**mAP**$_{[50]}$	**mAP**$_{[75]}$	**FPS**	**feature maps**
DTX	0.927	0.581	36	[3,4]
DTX	0.930	0.590	36	[2,4]
Model*	**mAP**$_{[50]}$	**mAP**$_{[75]}$	**FPS**	
DTX	0.926	0.557	**36**	
DTX + MS	0.939	0.561	32	
DTX + MS + CL	**0.945**	**0.572**	32	

* These experiments use the best hyper-params from grid search
DTX, Dense Transformer (our method); Contrastive Loss, CL
MS, Multi-scale backbone.

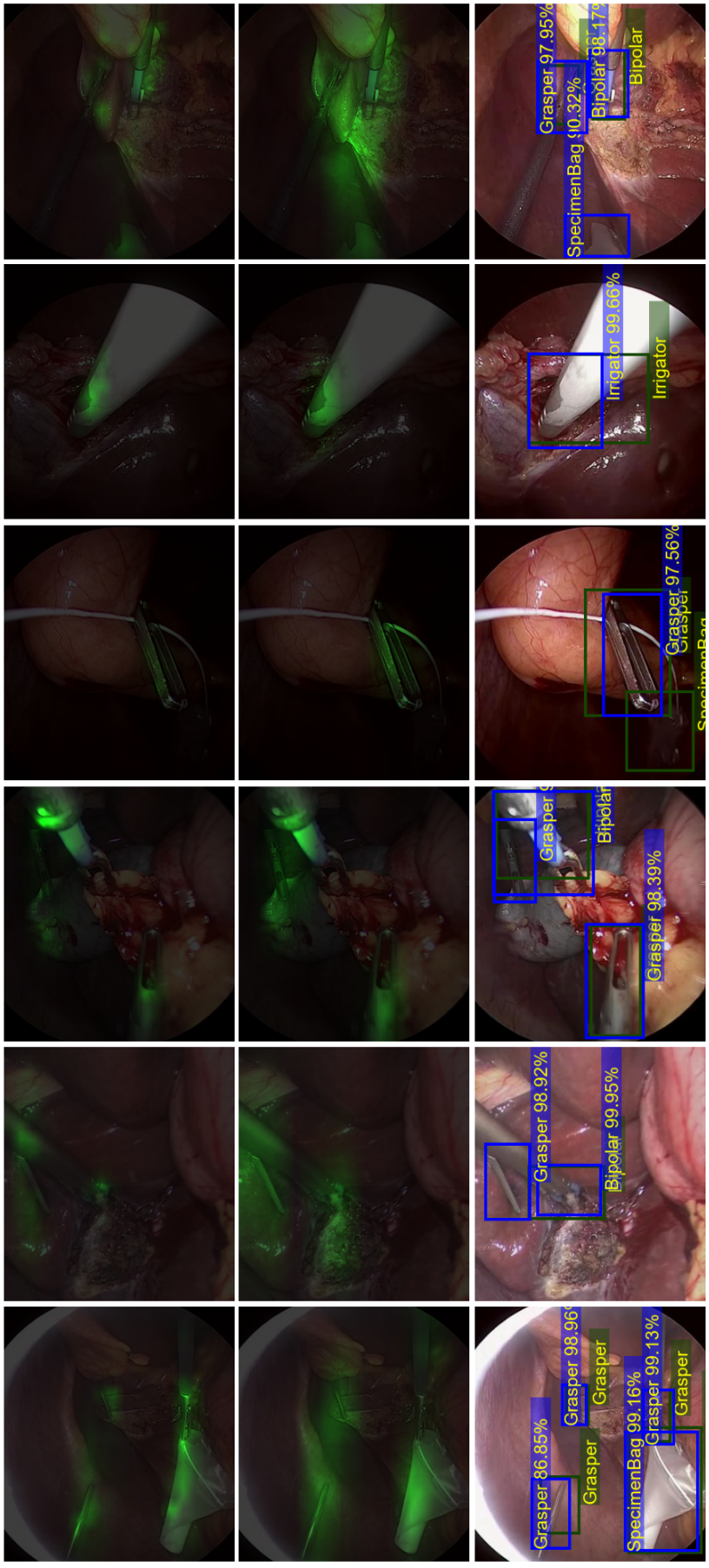


Figure A.1: **Attention maps for two scales**. An attention map for the relevant object representations in each image is provided. This map was divided into the different scales that were used in our final network.

Appendix B

Complementary material Ch. 4

B.1 Experiments on complementary datasets

To further assess the generalisation capabilities of the proposed framework, additional experiments were conducted on colonoscopic data. Laparoscopic scenarios often depict relatively consistent tissue regions across consecutive frames due to constrained camera motion. In contrast, colonoscopic sequences typically involve larger camera trajectories, larger viewpoint variation, stronger exposure changes, and increased geometric variability across the FoV. These characteristics introduce additional challenges for pose-informed reconstruction and volumetric optimisation. To test the robustness of our model under these conditions, experiments were performed on the SimCol dataset. These results were not included in the main manuscript due to the absence of ground-truth point correspondences in colonoscopic datasets, which prevents direct evaluation of tracking accuracy. However, two factors motivated this complementary study: (i) the availability of ground-truth depth, enabling quantitative reconstruction evaluation, and (ii) the emergence of real-time monocular pose estimation in robotic systems (e.g., ATLAS endoscopy), suggesting the practical applicability of pose-informed reconstruction even without stereo input.

B.1.1 Dataset

The SimCol (Synthetic Colonoscopy) dataset [114] is a synthetic dataset created for camera pose and depth estimation in colonoscopy. Built from real colon CT-derived anatomies to recreate realistic trajectories and imaging conditions. The dataset comprises more than 37,000 monocular RGB frames, together with camera intrinsics, depth maps, and pose annotations, across three anatomical models. For each anatomical model, the validation subset consists of three independent sequences.

B.1.2 Implementation remarks

The model was trained using the provided camera poses and evaluated using reconstruction metrics computed against the ground-truth depth maps. When estimating depth information to guide the optimisation process, the formulation of target points presented in Chapter 4.2.1 was followed, replacing stereo pairs with consecutive monocular frames. In addition to the indetermination condition described in Chapter 4, only triangulated points within a threshold distance of $5.5\,cm$ were considered valid for guiding the ray density. This constraint prevents noisy supervision from distant surfaces that appear poorly illuminated and textureless, and focuses optimisation on geometrically reliable regions close to the camera. To extract an explicit reconstruction from the InvNeRF representation, the same depth-validity mask was used to select points for tracking and to recover the 3D structure as an evolving mesh. The first frame initialises the reconstruction using a 2D grid of queried image-plane points. In subsequent frames, new points are introduced whenever large regions on the image plane lack tracked correspondences, enabling progressive growth of the reconstructed surface.

B.1.3 Results

Quantitative results are summarised in Table B.1. Despite being designed primarily for deformable laparoscopic scenarios, InvNeRF demonstrates competitive performance on colonoscopic sequences characterised by larger viewpoint changes and stronger photometric variation. In terms of absolute error (RMSE), InvNeRF achieves 0.642, comparable to Lite-Mono (0.640) and within 4.6% of the best-performing method (GLB+VAE, 0.614). Although its RMSE_{log} is higher than purely depth-supervised baselines, InvNeRF achieves the highest threshold accuracy across all δ metrics. This behaviour is consistent with the volumetric formulation of InvNeRF: optimisation prioritises geometric consistency along camera rays rather than direct pixel-wise depth regression. As a result, the representation favours spatial coherence and accurate localisation of surfaces, even under strong viewpoint changes.

Qualitative results in Figure B.1 further support this observation. Although the rendered RGB images do not closely match the ground-truth frames, the predicted depth maps preserve the global structural continuity of the colon surface. Importantly, the density-based rendering does not collapse in regions of exposure variation, indicating stable optimisation behaviour under challenging illumination, and the surface seems to be accurately tracked.

Figure B.2 shows the progressive extraction of an explicit 3D mesh from the im-

Table B.1: **Quantitative results for SimCol dataset.**

Model	rmse	rmse_{log}	δ_1	δ_1	δ_1
Monodepth2	0.709	0.200	0.837	0.946	0.979
MonoViT	0.715	0.157	0.863	0.952	0.952
Lite-Mono	0.640	0.144	0.886	0.969	0.990
Depth Pro	0.708	0.211	0.734	0.949	0.988
GLB+VAE	**0.614**	**0.126**	0.937	0.983	0.996
InvNeRF (ours)	0.642	0.359	**0.957**	**0.992**	**0.999**

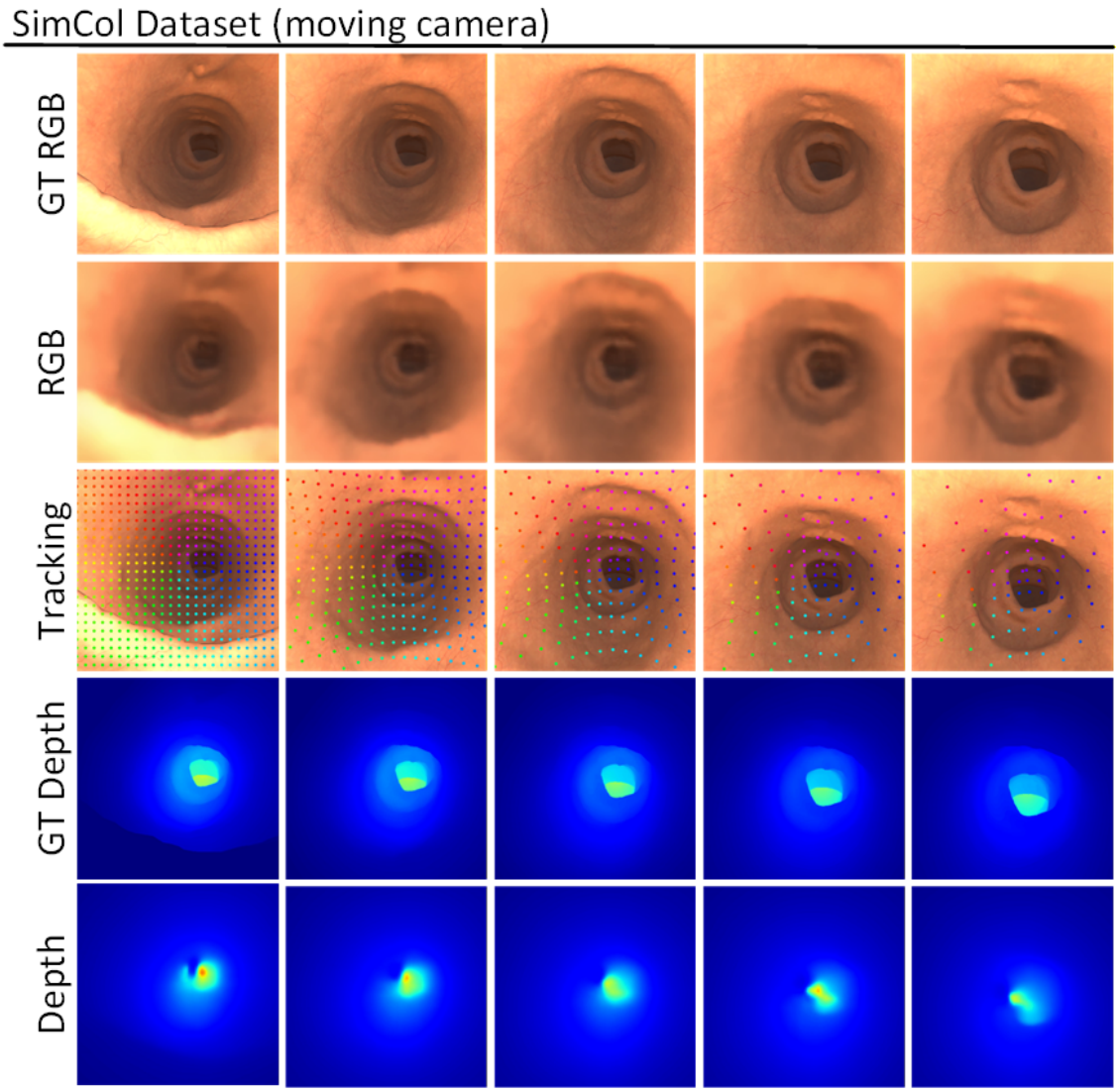


Figure B.1: **Qualitative results over the SimCol dataset.** Results over the initial frames of a validation video. The first row shows the GT RGB frames, followed by the rendered images and then tracked points (a grid is queried on on the first frame). The fourth row shows the GT depth, followed by the rendered depth.

plicit representation. The first row illustrates tracked points across frames, while the second row presents the growing surface reconstruction. The mesh exhibits coher-

ent surface evolution and maintains geometric consistency as new regions become visible. This demonstrates that the learned representation encodes a meaningful spatial structure rather than frame-wise independent depth predictions. Overall, these results provide preliminary evidence that the proposed framework could be applied beyond laparoscopic robotic settings and remains stable under the increased viewpoint variability typical of colonoscopy. Although tracking accuracy cannot be quantitatively evaluated due to the absence of ground-truth correspondences, the consistency of reconstruction metrics and the qualitative stability of the extracted mesh suggest that the architecture could scale to broader endoscopic applications.

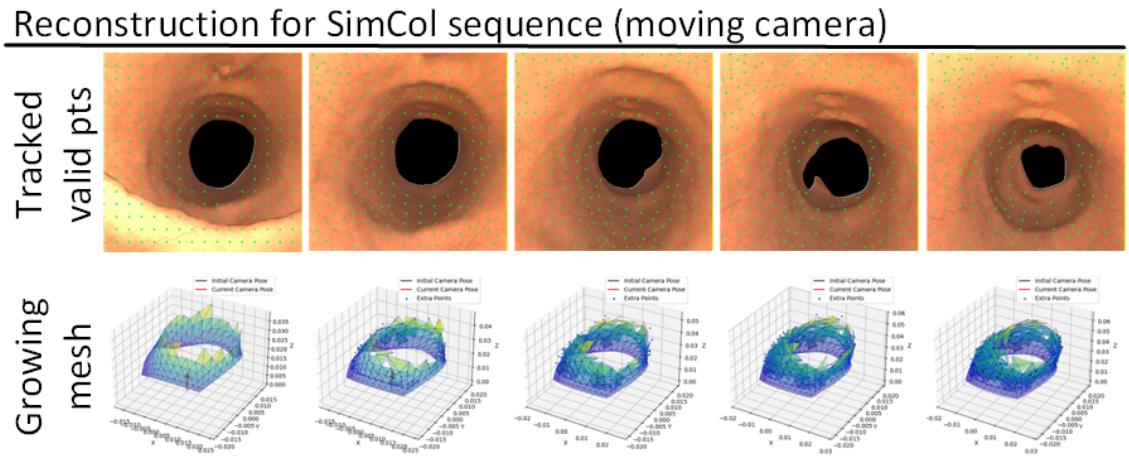


Figure B.2: **Reconstruction from InvNeRF representation** The first row contains valid tracked points along a sequence of frames, and the second row shows the growing 3D mesh of the colon (explicit representation extracted from the InvNeRF representation).