UNIVERSITY OF CLERMONT AUVERGNE

EnCoV-Institute Pascal

Engineering Sciences Doctoral School

P H D T H E S I S

to obtain the title of

PhD in Computer Science

Defended by

Rasoul SHARIFIAN

Contributions to 3D Reconstruction and Deformable Registration from Images with Application to Minimally Invasive Surgery

PhD director:	Adrien BARTOLI	–	Université Clermont Auvergne, CHU Clermont-Ferrand
Co-supervisor:	Navid RABBANI	–	CHU Clermont-Ferrand

defended on July 17, 2026

Jury Members :

Reviewers:	Tom VERCAUTEREN	-	King's College London
	Stamatia GIANNAROU	-	Imperial College London
President:	Jean-Sébastien FRANCO	-	Inria, Université Grenoble Alpes
Examiner:	Bertrand LE ROY	-	CHU Saint-Etienne

Abstract: Minimally-Invasive Surgery (MIS) has transformed surgical practice by enabling interventions through small incisions while improving postoperative recovery. However, vision-based surgical guidance in MIS remains highly challenging, as surgical scenes violate many assumptions of conventional computer vision, affecting the entire vision pipeline from geometric reconstruction to deformable registration. To address these challenges, this thesis investigates three complementary research axes: *(i)* the recovery of geometric structure from images under challenging MIS conditions and uncertain camera geometry, *(ii)* the computation of reliable correspondences under strong viewpoint and perspective distortions, and *(iii)* the registration of anatomical structures in the presence of complex deformations and topology changes. Together, these contributions aim to improve the robustness and reliability of vision-based surgical guidance systems.

The first part of the thesis addresses the problem of 2.5D reconstruction in MIS. We investigate two major approaches for this problem: the Monocular Single-shot Depth Estimation (MoSDE) and stereoscopic reconstruction. MoSDE is attractive in MIS because it only requires a standard monocular laparoscope. In addition, recent learning-based methods have shown promising performance, particularly in urban environments. However, their reliability in MIS remains limited by a severe domain shift. Furthermore, current evaluation protocols do not sufficiently assess robustness under such conditions. To address these limitations, this thesis introduces the Robust Depth Estimation in MIS (RoDEM) benchmark, a comprehensive evaluation framework for robust monocular depth estimation in MIS. It is built upon a new ex-vivo dataset comprising laparoscopic recordings of sheep organs with structured-light depth ground truth, across diverse common surgical perturbation conditions. In addition, the benchmark introduces evaluation protocols specifically designed to jointly assess reconstruction accuracy and robustness to perturbations.

The thesis then investigates stereoscopic reconstruction under uncertain camera geometry. Although stereo methods can recover metric depth under calibrated conditions, their performance strongly depends on accurate calibration and rectification. Calibration, however, is impractical to perform routinely in operating room conditions and may vary during surgery due to zoom, focus, or mechanical adjustments, making standard stereo pipelines unreliable in practice. To address the problem of uncalibrated stereo in MIS, we leverage large-scale non-medical synthetic stereo datasets. A major limitation of these datasets is that they are generated under ideal conditions, with perfectly rectified images and centred principal points. Moreover, they are typically released without full 3D scene information, preventing re-rendering. To overcome these limitations, we introduce Camera Augmentation Training Strategy (CATS), a framework for robust stereo matching in MIS under uncalibrated conditions. Specifically, we first introduce camera augmentation, which generates new training samples by perturbing camera orientation and intrinsic parameters, without requiring access to scene geometry or depth. Next, we use this

strategy to transform idealised stereo datasets into data representative of uncalibrated, yet near-rectified, stereo laparoscope configurations. We demonstrate that CATS enables standard stereo architectures, including RAFTStereo and IGEV++, to be adapted to uncalibrated settings, achieving reliable 2.5D reconstruction in MIS without task-specific retraining on limited medical data.

The second part of the thesis focuses on robust correspondence estimation under strong perspective distortions. Keypoint matching is a central component of numerous computer vision pipelines, however, its reliability significantly degrades under large viewpoint changes. In such situations, perspective projection induces complex non-isotropic distortions that cannot be approximated well by local affine transformations, thereby violating the assumptions underlying most standard keypoint detectors and descriptors. Moreover, assumptions such as scene planarity or developability are rarely satisfied in surgical environments. To address these limitations, this thesis introduces Conformal Perspective Correction (CPC), a geometric image-warping framework based on conformal surface flattening. Using the underlying 3D geometry of the observed surface, CPC constructs a locally fronto-parallel image representation over the entire surface. Although such an image does not correspond to a physically attainable camera view, it significantly reduces perspective distortions while preserving local shape and explicitly controlling image resampling. Experimental results demonstrate that CPC substantially improves keypoint matching under strong viewpoint variations and benefits downstream tasks including tracking and reconstruction on complex non-planar anatomical surfaces.

The final part of the thesis investigates deformable registration in MIS. This problem is particularly challenging because soft tissues undergo large non-linear deformations driven by physiological processes, surgical manipulation, and interactions with instruments. Furthermore, the observations available intraoperatively are sparse, while the true deformation-driving quantities, such as external forces and boundary conditions, remain unknown, making the registration problem inherently underconstrained. An additional challenge arises from topology changes caused by incision and resection. Conventional deformable registration methods typically rely on mesh-based representations with fixed connectivity, implicitly assuming topology preservation and therefore struggling to model discontinuities and often generating unrealistic stretching near incision regions. To address these limitations, this thesis introduces Registration by Optimisation over Differentiable Simulation (RODS), a physics-based framework formulated as an inverse simulation problem. Instead of directly estimating a deformation field, the method optimises the initial conditions of a differentiable simulation such that its forward evolution matches the observed intraoperative data, thereby constraining the solution space to physically plausible deformations. To naturally accommodate topology changes, the framework adopts a mesh-free particle representation based on the Material Point Method (MPM), resulting in the RODS-MPM formulation. By representing tissues as interacting material particles without fixed connectivity, the method allows separation and discontinuities to emerge naturally without explicit remeshing. Experimental re-

sults demonstrate that the proposed framework successfully models incision-induced topology changes and achieving superior Target Registration Error (TRE) performance compared with baseline methods.

Keywords: 3D Reconstruction, Keypoint Matching, Deformable Registration, Perspective Correction, Topology Changes, Minimally Invasive Surgery

Résumé :

La chirurgie mini-invasive (MIS) a profondément transformé la pratique chirurgicale en permettant des interventions à travers de petites incisions tout en améliorant la récupération postopératoire. Cependant, le guidage chirurgical basé sur la vision en MIS demeure particulièrement difficile, car les scènes chirurgicales violent de nombreuses hypothèses des méthodes conventionnelles de vision par ordinateur, affectant l'ensemble de la chaîne de traitement allant de la reconstruction géométrique au recalage déformable. Afin de répondre à ces défis, cette thèse étudie trois axes de recherche complémentaires : *(i)* la reconstruction de structure géométrique à partir d'images dans des conditions difficiles de MIS et avec une géométrie caméra incertaine, *(ii)* le calcul de correspondances robustes sous de fortes distorsions de perspective et changements de point de vue, et *(iii)* le recalage de structures anatomiques en présence de déformations complexes et de changements de topologie. Ensemble, ces contributions visent à améliorer la robustesse et la fiabilité des systèmes de guidage chirurgical basés sur la vision.

La première partie de cette thèse traite du problème de la reconstruction 2.5D en MIS. Deux grandes approches sont étudiées : l'estimation monoculaire de profondeur en une seule image (MoSDE) et la reconstruction stéréoscopique. Les méthodes MoSDE sont particulièrement attractives en MIS, car elles ne nécessitent qu'un laparoscope monoculaire standard. De plus, les approches récentes basées sur l'apprentissage profond ont montré des performances prometteuses, notamment dans des environnements urbains. Toutefois, leur fiabilité en MIS reste limitée en raison d'un important décalage de domaine. Par ailleurs, les protocoles d'évaluation actuels ne permettent pas d'évaluer suffisamment leur robustesse dans de telles conditions.

Afin de pallier ces limitations, cette thèse introduit le benchmark RoDEM, un cadre d'évaluation complet dédié à l'estimation robuste de profondeur monoculaire en MIS. Celui-ci repose sur un nouveau jeu de données ex vivo constitué d'enregistrements laparoscopiques d'organes ovins avec vérité terrain de profondeur obtenue par lumière structurée, couvrant diverses perturbations couramment rencontrées en chirurgie. Le benchmark propose également des protocoles d'évaluation spécifiquement conçus pour évaluer conjointement la précision de reconstruction et la robustesse aux perturbations.

La thèse étudie ensuite la reconstruction stéréoscopique sous géométrie caméra incertaine. Bien que les méthodes stéréoscopiques permettent de reconstruire une profondeur métrique dans des conditions calibrées, leurs performances dépendent fortement d'une calibration et d'une rectification précises. Or, la calibration est difficile à réaliser de manière routinière en salle d'opération et peut varier au cours de l'intervention en raison des ajustements de zoom, de focus ou de paramètres mécaniques, rendant les pipelines stéréoscopiques standards peu fiables en pratique. Pour répondre au problème de la stéréoscopie non calibrée en MIS, nous exploitons

de grands jeux de données stéréoscopiques synthétiques non médicaux. Toutefois, ces jeux de données sont généralement générés dans des conditions idéales, avec des images parfaitement rectifiées et des points principaux centrés. De plus, ils sont souvent diffusés sans les informations géométriques 3D complètes, empêchant tout nouveau rendu des scènes.

Afin de dépasser ces limitations, nous introduisons CATS, un cadre dédié à l'appariement stéréoscopique robuste en MIS dans des conditions non calibrées. Nous proposons tout d'abord une stratégie de camera augmentation générant de nouveaux échantillons d'apprentissage par perturbation de l'orientation des caméras et de leurs paramètres intrinsèques, sans nécessiter l'accès à la géométrie 3D des scènes ni aux cartes de profondeur. Cette stratégie permet ensuite de transformer des jeux de données stéréoscopiques idéalisés en données représentatives de configurations laparoscopiques non calibrées mais quasi rectifiées. Nous montrons que CATS permet d'adapter des architectures stéréoscopiques standards, telles que RAFTStereo et IGEV++, à des configurations non calibrées, afin d'obtenir des reconstructions 2.5D fiables en MIS sans nécessiter de réentraînement spécifique sur des données médicales limitées.

La deuxième partie de cette thèse s'intéresse à l'estimation robuste de correspondances sous fortes distorsions de perspective. L'appariement de points clés constitue un composant central de nombreux pipelines de vision par ordinateur ; cependant, sa fiabilité se dégrade fortement sous de grands changements de point de vue. Dans ces situations, la projection en perspective induit des distorsions complexes et non isotropes qui ne peuvent être correctement approximées par des transformations affines locales, violant ainsi les hypothèses sous-jacentes à la plupart des détecteurs et descripteurs de points clés standards. De plus, des hypothèses telles que la planarité ou la développabilité des surfaces sont rarement satisfaites dans les environnements chirurgicaux.

Pour répondre à ces limitations, cette thèse introduit CPC, un cadre géométrique de transformation d'image fondé sur l'aplatissement conforme de surfaces. À partir de la géométrie 3D de la surface observée, CPC construit une représentation d'image localement fronto-parallèle sur l'ensemble de la surface. Bien qu'une telle image ne corresponde pas à une vue physiquement réalisable par une caméra réelle, elle réduit significativement les distorsions de perspective tout en préservant localement les formes et en contrôlant explicitement le rééchantillonnage de l'image. Les résultats expérimentaux montrent que CPC améliore considérablement l'appariement de points clés sous de fortes variations de point de vue et bénéficie à des tâches en aval telles que le suivi et la reconstruction sur des surfaces anatomiques complexes et non planes.

La dernière partie de cette thèse porte sur le recalage déformable en MIS. Ce problème est particulièrement difficile car les tissus mous subissent d'importantes déformations non linéaires dues aux processus physiologiques, aux manipulations chirurgicales et aux interactions avec les instruments. De plus, les observations disponibles en peropératoire sont rares, tandis que les véritables quantités gouver-

nant les déformations, telles que les forces externes et les conditions aux limites, demeurent inconnues, rendant le problème intrinsèquement sous-contraint. Une difficulté supplémentaire provient des changements de topologie induits par les incisions et les résections. Les méthodes classiques de recalage déformable reposent généralement sur des représentations maillées à connectivité fixe, supposant implicitement la préservation de la topologie ; elles peinent ainsi à modéliser les discontinuités et génèrent souvent des étirements irréalistes à proximité des zones d'incision.

Afin de surmonter ces limitations, cette thèse introduit RODS, un cadre physique formulé comme un problème de simulation inverse. Plutôt que d'estimer directement un champ de déformation, la méthode optimise les conditions initiales d'une simulation différentiable de manière à ce que son évolution reproduise les observations peropératoires, contraignant ainsi l'espace des solutions à des déformations physiquement plausibles. Pour prendre naturellement en compte les changements de topologie, le cadre adopte une représentation particulaire sans maillage fondée sur la MPM, conduisant à la formulation RODS-MPM. En représentant les tissus comme des particules matérielles en interaction sans connectivité fixe, la méthode permet l'apparition naturelle de séparations et de discontinuités sans nécessiter de remaillage explicite. Les résultats expérimentaux montrent que le cadre proposé modélise avec succès les changements de topologie induits par les incisions tout en obtenant des performances de TRE supérieures à celles des méthodes de référence.

Mots-clés : Reconstruction 3D, Appariement de points clés, Recalage déformable, Correction de perspective, Changements de topologie, Chirurgie mini-invasive.

To Mina, my parents, and my siblings,
your unconditional love has been a constant light throughout my life.

Acknowledgments

I would like to thank my father, Mohammad Ali; my mother, Ashraf; and my siblings for a level of support that simply cannot be quantified. My deepest thanks go to my life companion, Mina, for always standing by my side through both difficult and joyful moments. I truly cherish the journey we have shared. I also wish to honour the memory of my brother Mohammad, whose unexpected loss reminded us once again how little time we have to love one another.

I would like to express my deepest gratitude to Adrien Bartoli. Adrien, thank you for your constant guidance, for listening to my ideas, and for pushing me to think beyond conventional boundaries. You have encouraged me to approach difficult problems with curiosity and boldness, and I am truly glad and grateful for every moment of the journey we have shared. Your trust, patience, and vision have fundamentally shaped how I think about research.

My special thanks go to Navid Rabbani. Navid, I am grateful for everything. You were not only a supervisor but also a constant source of support in many aspects of my life. Mina and I will always remember with gratitude how you and Neda welcomed us at the Salins bus station when we first arrived and how you have supported us throughout these years.

I am grateful to SurgAR for giving me this opportunity and for fostering such a motivating and inspiring environment throughout my CIFRE thesis. I would like to express my sincere gratitude to the entire team. This experience would not have been possible without your support.

I would also like to thank my colleagues at the EnCoV laboratory and the hospital, who became much more than collaborators and made this journey not only professionally rewarding but also personally meaningful.

I am also grateful to my former supervisors, Saeed Sadri and Behzad Nazari, who played important roles at different stages of my academic and personal development. Your guidance encouraged me to think more broadly about my life and future.

List of Publications

Journal articles

Camera Augmentation: Enabling Uncalibrated Stereo Matching of Minimally-Invasive Surgery Images by Training from the Wealth of Public Synthetic Image Datasets
R. Sharifian, N. Rabbani, Y. Zhang, and A. Bartoli
In *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, accepted April 2026.

SurgIPC: A Convex Image Perspective Correction Method to Boost Surgical Keypoint Matching
R. Sharifian and A. Bartoli
In *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, 20:1491–1502, June 2025.

The RoDEM Benchmark: Evaluating the Robustness of Monocular Single-shot Depth Estimation Methods in Minimally-Invasive Surgery
R. Sharifian, N. Rabbani, and A. Bartoli
In *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, 20:1215–1229, April 2025.

Automatic Smoke Analysis in Minimally Invasive Surgery by Image-based Machine Learning
R. Sharifian, H. Mendonça Abrão, S. Madad-Zadeh, C. Sève-d'Erceville, P. Chauvet, N. Bourdel, M. Canis, and A. Bartoli
In *Journal of Surgical Research*, 296:325–336, April 2024.

Kidney Tracking for Live Augmented Reality in Stereoscopic Mini-invasive Partial Nephrectomy
K. Chandelon, R. Sharifian, S. Marchand, A. Khaddad, N. Bourdel, N. Mottet, J.-C. Bernhard, and A. Bartoli
In *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, 11(4):1251–1260, June 2023.

International conferences

Camera Augmentation: Enabling Uncalibrated Stereo Matching of Minimally-Invasive Surgery Images by Training from the Wealth of Public Synthetic Image Datasets
R. Sharifian, N. Rabbani, Y. Zhang, and A. Bartoli
In *International Conference on Information Processing in Computer-Assisted Interventions (IPCAI)*, 2026.

SurgIPC: A Convex Image Perspective Correction Method to Boost Surgical Keypoint Matching
R. Sharifian and A. Bartoli
In *International Conference on Information Processing in Computer-Assisted Interventions (IPCAI)*, 2025.

The RoDEM Benchmark: Evaluating the Robustness of Monocular Single-shot Depth Estimation Methods in Minimally-Invasive Surgery
R. Sharifian, N. Rabbani, and A. Bartoli
In *International Conference on Information Processing in Computer-Assisted Interventions (IPCAI)*, 2025.

Workshops

Guiding Breast-Conservative Surgery by Augmented Reality from Preoperative MRI: Initial System Design and Retrospective Trials
R. Sharifian, S. Madad-Zadeh, N. Bourdel, A. Giro, W. Marraoui, C. Pomel, and A. Bartoli
In *DEEP-BREATH Workshop, MICCAI*, September 2024.

Preprints and submissions

Registration by Optimisation over Differentiable Simulation: Topology-adaptive Deformable Registration with MPM
R. Sharifian, N. Rabbani, and A. Bartoli
Under review, submitted to *ECCV*, 2026.

Contents

List of Acronyms

Chapter 1

Introduction and Thesis Contributions

Contents

1.1 Context

This thesis was conducted as part of a collaboration between DIA2M[1]/DRCI[2]/CHU de Clermont-Ferrand, and the SurgAR[3] company. It was carried out within the framework of a CIFRE[4] PhD fellowship (No. 2022/1015), co-funded by ANRT[5]. The objective of this partnership has been to develop visual computing solutions for Augmented Reality (AR) in surgical applications across multiple anatomical contexts, including minimally-invasive procedures involving the uterus, liver, and kidneys, as well as open surgery scenarios such as breast-conserving interventions.

This chapter first introduces the clinical context of the thesis, namely MIS, and describes its principles and associated challenges. It then presents the role of intraoperative guidance, with a particular focus on AR. Finally, the chapter outlines the motivation, research questions, and main contributions of the thesis toward enabling reliable AR in MIS.

1.2 Minimally-Invasive Surgery

MIS refers to surgical procedures performed through small incisions using specialised instruments and imaging systems, as opposed to traditional open approaches that require large operative exposures [Robinson and Stiegmann, 2004]. It fundamentally alters how the surgeon perceives and interacts with the anatomy, introducing specific visual and operational constraints. As such, MIS relies on a specific setup with its own characteristics and constraints.

1.2.1 General Points

In a typical MIS setup, access to the internal anatomy is achieved by inserting trocars, which are narrow tubular access ports, through the patient's body. An endoscopic camera system, typically a laparoscope, is introduced through one of these ports, while instruments, such as graspers, forceps, or scissors, are inserted through others to manipulate tissue and perform surgical actions, as illustrated in figure 1.1.

This configuration fundamentally changes the way the surgeon interacts with the anatomy. Unlike open surgery, where the surgeon directly observes and touches the organs, MIS relies on indirect visualisation through a video feed displayed on external monitors. The laparoscope provides a view of the surgical scene, often with

[1] Département Intelligence Artificielle et Applications Médicales

[2] Direction de la recherche clinique et de l'innovation, URL: https://www.chu-clermontferrand.fr/liste-services/direction-de-la-recherche-clinique-et-de-linnovation-drci

[3] Surgical Augmented Reality company, 22 Allée Alan Turing, 63000 Clermont-Ferrand, France. URL: https://surgar-surgery.com

[4] Convention Industrielle de Formation par la Recherche

[5] French National Agency for Research and Technology, URL: https://www.anrt.asso.fr/fr/le-dispositif-cifre-7844

Figure 1.1: Photograph of a standard MIS operating room, illustrating trocar placement, the laparoscope, surgical instruments, and display monitors, with surgeons standing around the patient and performing the procedure via the display screens. Source: B. Braun [B. Braun, 2021].

an embedded light source, but the perception of depth and spatial relationships must be inferred from the image. The surgeon operates while looking at a screen, coordinating hand movements with visual feedback that is spatially decoupled from the physical action.

MIS can be performed either using conventional laparoscopic techniques or with the assistance of robotic platforms [Brody and Richards, 2014]. Robotic systems typically consist of a surgeon console, one or more instrument arms, and an endoscopic imaging system, providing enhanced dexterity, motion scaling, stability, and improved ergonomics. They are often coupled with stereoscopic imaging to facilitate depth perception. Regardless of the platform, the surgical workflow fundamentally relies on an imaging system, where the camera may be monocular or stereoscopic. In monocular systems, the laparoscope is equipped with a single optical channel and image sensor. These systems remain prevalent in clinical practice but do not provide explicit depth information, which must instead be inferred from visual cues. By contrast, stereoscopic systems are configured with two optical channels capturing left and right views, enabling depth estimation through binocular disparity. However, they rely on accurate calibration and introduce additional hardware and computational complexity. In this thesis, we consider both configurations, independently of whether the procedure is robot-assisted or not.

A typical MIS operating room therefore consists of the patient, the laparoscopic equipment, and one or more display screens showing the endoscopic view. The

surgeon and assistants stand around the patient while focusing on these monitors, manipulating instruments that pivot around the entry points in the body, as illustrated in figure 1.1. This configuration results in a fundamentally different mode of surgical interaction compared to open surgery, in which the surgeon directly couples visual perception with hand movements.

MIS has been widely adopted across many surgical domains, including abdominal, thoracic, gynaecological, and urological procedures. In gynaecology [Chittawar et al., 2014], for instance, laparoscopic interventions are routinely performed on the uterus for procedures such as myomectomy or hysterectomy. In hepatobiliary surgery [Thiruchelvam et al., 2021], MIS is increasingly used for liver resections, where precise localisation of tumours and preservation of vascular structures are critical. In urological surgery, minimally-invasive approaches are standard for kidney surgeries, such as partial nephrectomy [Rashid et al., 2008], where tumorous tissue must be removed while preserving as much healthy parenchyma as possible.

1.2.2 Advantages and Drawbacks of MIS over Open Surgery

The clinical benefits of MIS are well established [Chen et al., 2017, Okunrintemi et al., 2016, Siddaiah-Subramanya et al., 2017]. By reducing incision size, it minimises tissue trauma, intraoperative blood loss, and postoperative pain. Patients benefit from shorter hospital stays, faster recovery, and improved cosmetic outcomes. These advantages have driven the widespread adoption of MIS and continue to motivate its extension to increasingly complex procedures.

However, these benefits come at the cost of significantly constrained perception and interaction. First, the field of view is limited and restricted to the region captured by the endoscopic camera, reducing global awareness of the surgical scene.

Second, the visualisation is inherently indirect. In monocular systems, the surgeon observes a two-dimensional projection of a three-dimensional scene. Even in stereoscopic systems, depth perception remains less intuitive than in direct vision and may be affected by system limitations or calibration issues. Third, tactile feedback is substantially reduced. The use of long instruments attenuates force feedback, making it more difficult to assess tissue properties such as stiffness and tension. As a result, the surgeon must predominantly rely on visual cues to interpret the state of the tissue. Fourth, hand–eye coordination is altered. Instrument movements are constrained by the trocar insertion points, creating a fulcrum effect [Nisky et al., 2012] that inverts or modifies motion. The surgeon must therefore perform actions in a transformed coordinate system, while simultaneously interpreting visual feedback indirectly through monitors. In conventional laparoscopy, this results in non-intuitive motion and significantly complicates hand–eye coordination. Robot-assisted systems partially mitigate this effect through motion scaling and intuitive re-mapping of instrument movements, improving dexterity and ergonomics. Nevertheless, the interaction remains indirect and primarily vision-driven, imposing a substantial cognitive load and requiring significant expertise. Finally, access to

internal or occluded anatomical structures is limited. Critical elements such as tumours, vessels, or functional regions are often not directly visible in the endoscopic image. In procedures such as liver resection or partial nephrectomy, the surgeon must rely on preoperative imaging (e.g., Computed Tomography (CT) or Magnetic Resonance Imaging (MRI)) and mental reasoning to estimate the location of hidden structures. This misalignment between visible and underlying anatomy constitutes a major source of uncertainty in surgical decision-making.

These limitations lead to a central challenge in MIS: surgical decision-making must be performed under incomplete, indirect, and sometimes ambiguous spatial information. This is particularly critical in procedures involving complex anatomy or high-risk regions, where precise localisation and navigation are essential.

1.2.3 Intraoperative Guidance and Augmented Reality

To address a subset of these limitations, in particular those related to perception and spatial understanding, intraoperative guidance systems have been developed to enhance the surgeon's visual feedback and provide additional spatial context. These include, for example, intraoperative ultrasound, fluorescence imaging, and surgical navigation systems. Among these approaches, AR has emerged as a particularly promising paradigm. In the MIS context, AR aims to enrich the surgical view by superimposing patient-specific 3D information onto the endoscopic image. This enables the visualisation of otherwise invisible structures, such as subsurface tumours and critical blood vessels, as illustrated in figure 1.2.

AR can thus be described as a navigation framework structured into three main stages:

First, extraction of guiding information. This stage consists in extracting the relevant anatomical information, typically from preoperative CT or MRI data, and reconstructing patient-specific 3D models through segmentation.

Second, registration of preoperative models to intraoperative images. This stage aims to estimate the position of the preoperative guiding information within the intraoperative image. To this end, the deformation or the displacement

Figure 1.2: Schematic illustration of 3D–2D registration for AR in MIS. A preoperative 3D model is registered to an intraoperative laparoscopic image. The registered model is then projected onto the image plane to generate an augmented view, enabling the visualisation of subsurface structures.

field relating the preoperative 3D model to the intraoperative anatomy must be recovered. This is generally the most complex and critical step of the pipeline.

Third, visualisation of the registered models. This stage determines how the registered 3D information is projected and displayed within the surgical view. It directly affects perceptual guidance and involves design choices such as colour encoding, transparency, shading, and depth cues.

In practice, additional computer vision components are often integrated into this pipeline to improve robustness and usability. One important component is 2.5D reconstruction, which estimates the geometry of the observed scene from monocular or stereo images. This may be achieved through stereo matching, structure-from-motion, or monocular depth estimation methods. The resulting depth information provides geometric cues about the intraoperative scene and may support downstream tasks such as registration, visualisation, collision avoidance, and surgical navigation. In MIS, stereo reconstruction is particularly attractive as modern robotic and laparoscopic systems increasingly provide stereoscopic imaging.

Another important component is keypoint detection and matching. These methods establish correspondences between images by identifying repeatable visual features and matching them across viewpoints or time. Such correspondences can provide geometric constraints for camera pose estimation, tracking, reconstruction, and registration. However, reliable matching in MIS remains challenging due to tissue deformation, specularities, smoke, weak texture, and strong perspective distortions.

Finally, to maintain the AR overlay consistently over time, the system must continuously estimate the relative motion between the camera and the anatomy. This is achieved through tracking, which propagates the registration across successive video frames. Tracking may rely on keypoint matching, dense image alignment, or optical flow-based approaches. Robust tracking is essential for live AR, as registration performed only on isolated frames would rapidly become invalid due to camera motion, respiration, tissue deformation, or surgical manipulation.

Figure 1.3 illustrates the overview of a general vision-based surgical guidance pipeline including its main stages and optional components. When successfully achieved, AR provides the surgeon with a geometrically consistent augmented view of the scene, enabling intuitive interpretation of subsurface structures and support-

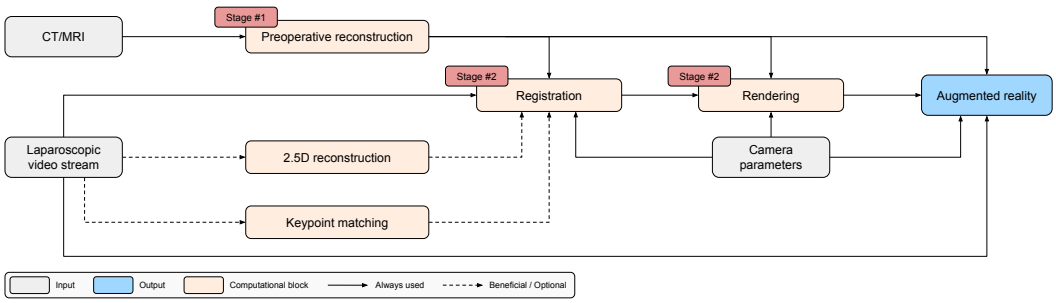


Figure 1.3: Overview of a general vision-based surgical guidance pipeline.

ing accurate and informed surgical actions.

1.3 Challenges in Vision-based Surgical Guidance

The conditions encountered in MIS differ significantly from those typically assumed in standard computer vision. As a result, methods that perform well in conventional computer-vision settings often degrade or fail when applied to surgical imagery. The challenges can be broadly categorised into three groups: those related to image formation, camera geometry, and scene and anatomical properties.

1.3.1 Image Formation Challenges

The visual appearance of surgical scenes is governed by specific imaging conditions that strongly deviate from classical assumptions in computer vision. In most endoscopic systems, the light source is co-located with the camera, resulting in highly directional, viewpoint-dependent illumination, as opposed to the common assumption of distant (light-at-infinity) sources [Modrzejewski et al., 2020]. This configuration produces strong specular reflections on moist tissue surfaces, thereby violating the Lambertian reflectance assumption commonly used in many vision algorithms [Maier-Hein et al., 2013]. In addition, the intraoperative environment introduces perturbations that degrade image quality. Smoke generated during cauterisation, motion blur caused by camera or tissue movement and occasional lens contamination can all obscure visual information. These factors reduce contrast, introduce noise, and may temporarily occlude relevant structures.

1.3.2 Camera Configuration and Viewpoint Challenges

Beyond appearance, the configuration of imaging systems in MIS introduces additional complexity. While stereo laparoscopes are increasingly adopted, monocular systems remain prevalent in clinical practice. In such setups, scene structure is not directly available and must be inferred from visual cues. Stereoscopic systems provide an alternative by enabling depth estimation; however, reliable calibration cannot always be assumed available [Sharifian et al., 2026]. In practice, calibration may degrade over time due to mechanical stress, temperature variations, or changes in camera settings such as zoom and focus. These factors affect both intrinsic and extrinsic parameters, thereby reducing the accuracy of triangulation and depth reconstruction. Furthermore, the laparoscope is frequently moved during the procedure to explore different regions of the anatomy. These viewpoint changes can be significant, especially in confined spaces where the camera operates close to the tissue. As a result, images of the same anatomical structure may exhibit strong perspective distortions, particularly on non-planar surfaces. These distortions make matching and tracking more challenging than in many conventional imaging settings.

1.3.3 Scene and Anatomy Challenges

The nature of the observed scene in MIS introduces some of the most significant challenges. Unlike rigid environments commonly considered in classical computer vision, surgical scenes are dominated by soft tissues that undergo complex deformations [Collins et al., 2020, Labrunie et al., 2023]. These deformations may be caused by physiological processes such as respiration, by interaction with surgical instruments, or by changes in patient positioning. More importantly, surgical actions such as cutting, incision, and resection can introduce large, non-smooth deformations and even topology changes [François et al., 2021, Paulus et al., 2017]. For example, during a liver resection or a partial nephrectomy, part of the organ is removed, altering its topological structure. Similarly, in uterine procedures, tissue manipulation may significantly change the shape and connectivity of anatomical structures. These phenomena violate the assumptions of many classical registration methods, which typically rely on smooth and topology-preserving transformations.

1.3.4 Implications for Vision-Based Surgical Guidance

Taken together, the above challenges highlight that surgical visual computing operates in conditions where both image formation and scene geometry deviate significantly from the assumptions underlying standard computer vision pipelines. Image formation is non-ideal, camera geometry may be uncertain, and the observed scene is dynamic, deformable, and subject to topology changes.

As a result, methods designed for conventional settings cannot be directly applied without adaptation. Instead, there is a need for approaches that explicitly account for the specific characteristics of MIS, including robustness to specific perturbations, flexibility in handling imperfect calibration, and the ability to model complex anatomical deformations. These considerations motivate the focus of this thesis, which addresses the challenges of scene reconstruction from surgical images, establishing reliable correspondences under difficult viewing conditions, and registering anatomical models in the presence of deformation and topology change.

1.4 Research Questions and Contributions

The challenges outlined in section 1.3 highlight that vision-based surgical guidance operates under conditions that fundamentally depart from the assumptions of conventional computer vision. These challenges are not isolated: they affect the entire pipeline, from low-level image understanding to high-level geometric reasoning and registration. Addressing them therefore requires a structured investigation of the key components that enable reliable spatial understanding in MIS.

1.4.1 Core Methodological Contributions

In this thesis, we organise our investigation along three complementary axes:

- the recovery of geometric structure from single-shot or uncalibrated stereo images under challenging MIS conditions and uncertain camera geometry;
- the computation of reliable correspondences under challenging viewpoint conditions;
- the registration of anatomical structures in the presence of complex deformations and topology changes.

These axes give rise to four Research Question (RQ), each addressed by a corresponding group of contributions.

RQ1 — *Are state-of-the-art learning-based monocular 2.5D reconstruction methods reliable in MIS under realistic surgical perturbations?*

MoSDE has achieved significant progress with learning-based approaches, yet its evaluation in surgical contexts remains limited. In particular, existing benchmarks often fail to capture the variability and perturbations encountered in real procedures, raising questions about the reliability and generalisability of these methods.

Contribution. This thesis introduces the RoDEM benchmark, designed to evaluate monocular depth estimation under realistic surgical conditions, including domain-specific perturbations such as smoke, specularities, and bleeding. This benchmark provides a systematic framework for assessing robustness and generalisation of MoSDE methods in MIS.

RQ2 — *How can stereo matching methods trained on accurately calibrated image pairs remain effective in MIS when intraoperative camera calibration is unavailable?*

While stereo imaging offers the potential for metric depth estimation, it typically relies on accurate camera calibration. In practice, however, operating room constraints may lead to imperfect or unavailable calibration, particularly in complex clinical environments. This raises the need for approaches that can estimate depth under such conditions, through exploiting the geometric properties of stereoscopic laparoscopes to provide meaningful depth perception.

Contribution. This thesis proposes the CATS, which enables stereo matching in uncalibrated or weakly calibrated settings by explicitly modelling variations in camera geometry during training. By leveraging synthetic datasets augmented through camera parameter perturbations, the proposed approach allows state-of-the-art stereo models to generalise to real MIS conditions without requiring calibration.

RQ3 — *How can strong perspective distortions be mitigated to improve image matching?*

Viewpoint changes in MIS often induce strong perspective effects. Due to the closeness of the camera to the organ, even small camera motions may result in

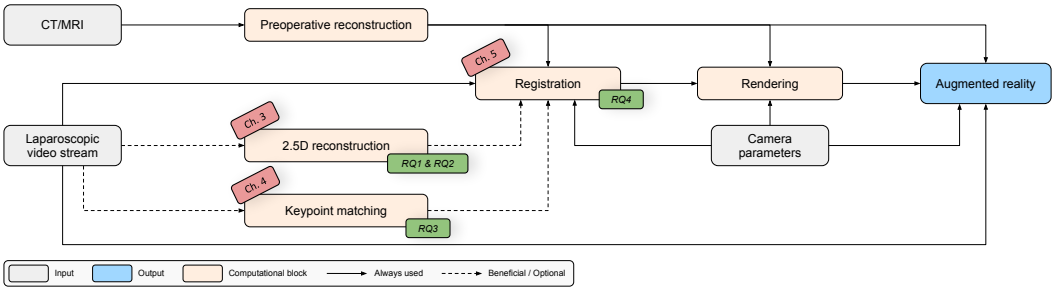


Figure 1.4: The positioning of the thesis contributions, research questions (RQ1–RQ4) and their corresponding chapters within a vision-based surgical guidance pipeline.

large perspective distortions. In addition, relative motion between the camera and deforming tissues leads to significant viewpoint variations. These distortions degrade the performance of standard keypoint matching pipelines and affect the stability of downstream tasks of tracking and reconstruction. Addressing this issue requires methods that explicitly account for the geometric effects of perspective.

Contribution. This thesis introduces a perspective-correction method based on conformal flattening, which reduces perspective distortions while preserving the underlying image signal. This approach improves correspondence estimation under challenging viewpoints and enables more stable matching in surgical images.

RQ4 — *How can deformable registration be achieved when the organ undergoes large deformations and topology changes?*

Conventional registration methods typically assume smooth and topology-preserving deformations. However, surgical procedures frequently involve cutting and resection, leading to discontinuities and structural changes. This raises the need for registration frameworks that can handle both large deformations and topology modifications.

Contribution. This thesis introduces a topology-adaptive deformable registration framework based on differentiable simulation. The proposed approach formulates registration as an optimisation problem over a differentiable physical simulation, enabling gradient-based estimation of physically plausible deformations. By leveraging the MPM, the framework naturally accommodates large deformations and topology changes, providing a flexible and realistic model of surgical scenarios.

Taken together, these four questions define a coherent research programme centred on the 3D reconstruction and registration of spatial information in MIS. They reflect a progression from estimating geometric structure, to establishing robust correspondences under challenging imaging conditions, and ultimately to achieving anatomically consistent registration in dynamic surgical scenes. This progression, together with the associated contributions, is summarised in figure 1.4, which provides a unified view of the MIS guidance pipeline and the positioning of this thesis.

This thesis addresses the research questions introduced in the previous section through a set of contributions that collectively aim to improve the reconstruction and registration computation blocks in the MIS pipeline. These contributions are organised into four groups, corresponding to the main challenges of 2.5D reconstruction, robust correspondence under strong viewpoint changes, and deformable registration, as well as complementary translational works.

1.4.2 Complementary Contributions

In addition to the core methodological contributions, this thesis includes two complementary works.

The first complementary contribution concerns AR for Breast Conserving Surgery (BCS), where patient-specific models are reconstructed from preoperative MRI acquired in supine positioning. This work highlights the importance of reconstruction and registration techniques in open surgery contexts and is presented in **Appendix A**.

The second complementary contribution addresses automatic smoke analysis in MIS. This work focuses on the development of image-based machine learning methods for quantifying the amount of surgical smoke and estimating its impact on smoke evacuation decisions. The work is presented in **Appendix B**.

In addition, **Appendix C** presents extended experimental results for the RoDEM dataset, providing additional quantitative and qualitative evaluations complementary to the results reported in the main manuscript.

1.5 Thesis Organisation

The remainder of this manuscript is organised in accordance with the research questions and contributions introduced in section 1.4, following the progression from geometric reconstruction to correspondence estimation and deformable registration. It should be noted that chapter 2 provides the general theoretical background. Detailed, problem-specific background and literature are presented within each contribution chapter.

Chapter 2 — Technical background and preliminaries This chapter introduces the theoretical and methodological foundations required for the rest of the thesis. It covers key concepts in MIS imaging, depth and stereo reconstruction, projective geometry, perspective distortion, conformal flattening, and deformable registration.

Chapter 3 — 2.5D reconstruction of the surgical scene from monocular methods or uncalibrated stereoscopes This chapter addresses the problem of 2.5D reconstruction of the scene from surgical images, corresponding to research questions RQ1 and RQ2. It presents the RoDEM benchmark for evaluating monocular depth estimation methods under realistic surgical conditions and introduces the

CATS framework for enabling stereo matching in uncalibrated or weakly calibrated settings.

Chapter 4 — Improving image keypoint matching under strong perspective distortions via conformal flattening This chapter focuses on the problem of establishing reliable correspondences under strong perspective distortions, addressing research question RQ3. It introduces the proposed perspective-correction based on conformal flattening designed to mitigate viewpoint-induced distortions and improve matching and tracking performance in MIS.

Chapter 5 — Topology-adaptive deformable registration by differentiable simulation This chapter addresses deformable registration under large deformation and topology change, corresponding to research question RQ4. It presents the RODS framework, which formulates registration as an optimisation problem over a differentiable physical simulation. The approach enables physically plausible registration of anatomical structures in surgical scenarios involving cutting and resection.

Chapter 6 — Conclusion and future work This chapter summarises the main findings of the thesis, discusses limitations, and outlines directions for future research.

Appendices — Complementary contributions and extended results These appendices present complementary contributions on AR for BCS and automatic smoke analysis in MIS, as well as extended results on the RoDEM dataset.

Chapter 2

Technical Background and Mathematical Preliminaries

Contents

2.1 Chapter Overview

The goal of this chapter is to present the preliminary mathematical models and tools used throughout this thesis. It first introduces the pinhole camera model, an image formation model, and stereo geometry, which provide the foundations for relating 3D scene structure and 2D images. In this thesis, we mostly focus on the geometric aspects of image formation, namely the mapping between 3D scene points and their 2D image projections. More generally, synthesising realistic images also requires a scene representation and additional components such as lighting and reflectance modelling. There exist advanced scene representation approaches well adapted to vision tasks, such as Neural Radiance Fields (NeRF) [Mildenhall et al., 2021] and 3D Gaussian Splatting (3DGS) [Kerbl et al., 2023], however they are not in the scope of this thesis. It then discusses 2.5D reconstruction from monocular and stereo images, including the associated geometric ambiguities. The chapter next discusses the hierarchy of geometric transformations including perspective projection, and

their geometric effects. These concepts are particularly relevant to the perspective correction framework developed in chapter 4. Additional literature review and methodological details specific to each topic are provided at the beginning of the corresponding chapters.

2.2 Geometry of Monocular and Stereo Camera

Camera models provide the base mathematical framework relating a three-dimensional scene to its two-dimensional image observations. We first introduce the pinhole camera model and the associated notions of intrinsic and extrinsic parameters. We then review stereo geometry and epipolar constraints, before introducing 2.5D reconstruction from monocular and stereo images.

2.2.1 Pinhole Camera Model

A camera maps a three-dimensional (3D) scene onto a two-dimensional (2D) image formed on its sensor. In computer vision, this process is commonly modelled by *central perspective projection*, whose canonical formulation is the *pinhole camera model*. This model provides a simplified geometric description of image formation by abstracting away the physical complexity of multi-lens systems and focusing instead on the projective relationship between 3D scene points and their 2D projections. Despite its simplicity, it has proven highly accurate at modelling a wide range of camera devices and makes numerous geometric vision tasks tractable.

In this model, the scene is observed from a single projection centre, referred to as the *optical centre*, which defines the origin of the camera coordinate system. All light rays pass through this point and intersect the retina, located at a focal distance f. The Z-axis defines the *optical axis*, corresponding to the viewing direction of the camera.

A 3D point expressed in the camera coordinate system, $\mathbf{P}_c = (X_c, Y_c, Z_c)^\top$, is projected onto the image plane at $\mathbf{p} = (u, v)$ according to a perspective mapping:

$$\begin{aligned} u &= f_x \frac{X_c}{Z_c} + c_x, \\ v &= f_y \frac{Y_c}{Z_c} + c_y, \end{aligned} \tag{2.1}$$

where f_x and f_y denote the focal length expressed in pixels along the horizontal and vertical image axes, respectively. Although the camera has a single physical focal length, two parameters are used because the conversion from metric coordinates to pixel coordinates may differ between the two image directions, due to possibly non-square pixels. The parameters (c_x, c_y) correspond to the principal point, defined as the intersection between the optical axis and the retina, also in pixels. This

projection can be expressed in homogeneous coordinates as:

$$\underbrace{\begin{bmatrix} u \\ v \\ 1 \end{bmatrix}}_{\text{Image coordinates}} = \underbrace{\alpha}_{\text{Scale factor}} \underbrace{\begin{bmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{K,\ \text{Intrinsic matrix}} \underbrace{\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix}}_{\text{Camera coordinates}} \tag{2.2}$$

where $\alpha \in \mathbb{R}$ is an arbitrary scale factor inherent to homogeneous representations. This reflects the projective nature of image formation: projections are invariant to global scaling of the 3D scene, leading to the loss of absolute depth and the well-known scale ambiguity in reconstruction (see section 2.3.3).

In the general case, 3D points are expressed in a world coordinate system. The mapping from a world point $\mathbf{P}_w$ to its image projection involves both *intrinsic* and *extrinsic* parameters. The intrinsic parameters are encoded in the calibration matrix $\mathbf{K}$, while the extrinsic parameters define the rigid transformation between the world and camera coordinate systems. This transformation is given by $\mathbf{P}_c = \mathbf{R}\mathbf{P}_w + \mathbf{t}$, where $\mathbf{R} \in SO(3)$ is a rotation matrix belonging to the Special Orthonormal group, i.e. the set of orthonormal 3×3 matrices with determinant equal to 1, and $\mathbf{t} \in \mathbb{R}^3$ is a translation vector. The perspective projection model can therefore be written as:

$$\tilde{\mathbf{p}} = \alpha \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix} \tilde{\mathbf{P}}_w, \tag{2.3}$$

where $\tilde{\mathbf{P}}_w$ and $\tilde{\mathbf{p}}$ denote the homogeneous representations of the 3D world point and its image projection, respectively.

While the pinhole model captures the ideal perspective geometry, real cameras exhibit deviations due to lens imperfections. These effects are typically modelled using non-linear distortion functions. A widely used formulation is the Brown–Conrady model [Brown, 1971], in which a 3D point is first projected using the pinhole model, and the resulting image coordinates (u, v) are then transformed by a distortion function $d : \mathbb{R}^2 \to \mathbb{R}^2$, yielding distorted coordinates. The distortion function depends on the radial distance $r^2 = u^2 + v^2$ and is parameterised by coefficients α_i, $i \in \{1, \ldots, 5\}$, which model radial and tangential effects. When these coefficients are known, typically through camera calibration, the distortion can be inverted. This allows the image to be interpreted under an idealised pinhole camera model.

In the context of endoscopy, and more specifically in MIS, the pinhole model generally provides a sufficiently accurate approximation of rigid endoscopic imaging systems [Yamaguchi et al., 2004, Barreto et al., 2009, Maier-Hein et al., 2013]. Although optical and mechanical factors may introduce deviations, endoscopes can typically be modelled as central cameras after calibration, enabling the use of standard projective geometry tools in surgical vision applications.

2.2.2 Stereo Camera Geometry and Epipolar Constraints

Stereo systems rely on the simultaneous acquisition of a left and a right image, and are fundamentally based on epipolar geometry. Epipolar geometry defines the projective constraints relating corresponding points between two views. A 3D point $\mathbf{P}$ in the scene projects onto two points $\mathbf{p}_L$ and $\mathbf{p}_R$, located respectively in the image planes of the left and right cameras. The position of $\mathbf{p}_L$ constrains the possible location of $\mathbf{p}_R$ to a line in the right image, called the epipolar line, denoted d_R. The epipolar relation can be formalised by the fundamental matrix F, a 3×3 matrix that encodes all epipolar constraints independently of the intrinsic parameters of the cameras. Representing the image points in homogeneous coordinates, the epipolar constraint is written as:

$$\tilde{\mathbf{p}}_R^\top \mathbf{F} \tilde{\mathbf{p}}_L = 0, \tag{2.4}$$

Exploiting this constraint reduces the correspondence search from two dimensions to one, along the epipolar lines. In the particular case where the cameras are rectified, that is, their optical axes are parallel and their image planes are coplanar and oriented consistently, the epipolar lines become horizontal. This condition is achieved by applying rectification homographies H_G and H_D to the left and right images, respectively. In this setting, disparity is defined as the difference between the horizontal coordinates of corresponding pixels:

$$d = u_L - u_R, \tag{2.5}$$

where u_L and u_R are the horizontal coordinates of the projections of point $\mathbf{P}$ in the left and right images.

Under the rectified stereo model, the depth Z of a point can be recovered through triangulation:

$$z = \frac{fB}{d}, \tag{2.6}$$

where f is the focal length (in pixels) and B is the baseline, i.e. the distance between the two camera centres.

Disparity estimation in a rectified stereo system can be performed using classical methods such as Block Matching [Hamzah and Ibrahim, 2016, Hirschmuller, 2007], which generally produce sparse disparity maps. More recent methods based on deep learning, such as Foundation Stereo [Wen et al., 2025], allow the estimation of dense disparity maps. Equation 2.6, suggests that, regardless of the method used to estimate depth, rigorous stereo calibration remains essential, especially in the case where the metric depth is needed. This involves estimating not only the intrinsic parameters of each camera but also the rigid transformation relating them.

A common approach consists of calibrating each camera individually using a monocular method, then estimating the extrinsic parameters through joint observa-

tion of a calibration pattern. In practice, however, variations in imaging conditions, such as changes in focal length or focus, may affect both intrinsic and extrinsic parameters, leading to inaccuracies in depth estimation. This is particularly relevant in minimally invasive surgery, where calibration may not be stable throughout the procedure.

2.3 2.5D Reconstruction from Images

In computer vision, the term *2.5D reconstruction* refers to the estimation of scene geometry in the form of a depth map, in which a scalar depth value is associated with each pixel of an image. Unlike full 3D reconstruction, which aims to recover a complete and viewpoint-independent representation of the scene, 2.5D reconstruction provides a view-dependent description expressed in the camera coordinate system. As such, the reconstructed geometry is defined only along the camera viewing rays and does not explicitly model occluded or non-visible regions.

Formally, the task is to estimate a dense depth map $D \in \mathbb{R}^{H \times W}$ from an input image $I \in \mathbb{R}^{H \times W \times 3}$, where each value $d_{i,j}$ represents the distance between the camera and the corresponding scene point.

2.3.1 Monocular 2.5D Reconstruction

In the monocular setting, depth is inferred from a single image. Since the projection process inherently removes one spatial dimension, recovering depth from a single view is an ill-posed problem: multiple 3D scenes can produce the same 2D image under perspective projection. This ambiguity is directly related to the projective nature of the camera model (the scale factor in equation 2.2), where points along a viewing ray project to the same image location.

As a consequence, monocular methods can recover depth only up to an unknown transformation. This ambiguity appears as a global scaling, meaning that depth is defined up to a multiplicative factor. In more general formulations, particularly those based on inverse depth or disparity, the ambiguity extends to an affine transformation, where predictions are defined up to both a scale and a shift. These different forms of ambiguity reflect a fundamental limitation of monocular 2.5D reconstruction: without additional information, such as known scale, camera motion, or scene constraints, metric depth cannot be uniquely determined.

2.3.2 Stereo 2.5D Reconstruction

In contrast to monocular methods, stereo reconstruction exploits two views to recover depth through geometric constraints. Given two images of the same scene acquired from different viewpoints, depth can be computed by establishing pixel correspondences and using the disparity–depth relationship introduced in section 2.2.2.

Under known camera calibration, stereo reconstruction enables the recovery of metric depth through triangulation. This provides a strong geometric advantage over monocular approaches, as depth is directly constrained by image observations rather than inferred from learned priors.

However, the accuracy of stereo reconstruction critically depends on several factors. First, reliable correspondence estimation is required, which is challenging in regions with low texture, specular reflections, or repetitive patterns. Second, accurate camera calibration and rectification are assumed known and done, which may not hold in practice, particularly in MIS, where camera parameters can vary during acquisition. Recent deep learning approaches have significantly improved the robustness and density of disparity estimation, enabling high-quality depth reconstruction even in challenging conditions [Wen et al., 2025, Psychogyios et al., 2022]. Nevertheless, these methods remain grounded in the geometric principles of stereo vision and are still sensitive to calibration inaccuracies and domain-specific artefacts.

2.3.3 Invariance in Depth and Disparity Representations

Depending on the formulation, depth and disparity predictions may be defined up to specific transformation groups, which leads to different notions of invariance. In particular, scale-invariant depth is defined up to a multiplicative factor, i.e., $z' = \alpha z$, while affine-invariant disparity is defined up to a scale and shift, $d' = \alpha d + \beta$. Such invariances naturally arise from the learning objectives used in monocular depth estimation methods [Birkl et al., 2023].

These invariances extend to geometric entities derived from depth. A reconstructed scale-invariant 3D point is defined up to a global scaling of the scene, whereas an affine-invariant disparity representation is defined up to an affine transformation on disparity maps. In practice, this means that reconstructed geometry preserves relative structure while lacking absolute metric consistency.

These invariance properties have direct implications for both the evaluation of monocular 2.5D reconstruction methods and the integration of their predictions into downstream geometric tasks. In particular, comparing predicted outputs against ground-truth depth requires accounting for the transformation group under which the predictions are defined. Several methods predict disparity rather than absolute depth, with outputs defined up to an unknown scale and shift. Consequently, accurate evaluation requires estimating the optimal scale and shift parameters to align the predicted disparity with the ground truth, prior to computing depth metrics.

Several alignment strategies have been proposed in the literature [Birkl et al., 2023]. In this thesis, we distinguish three alignment strategies: no-alignment, scale alignment (s-alignment), and scale-and-shift alignment (s&t-alignment).

No-alignment. This strategy directly evaluates the predicted depth values without any transformation and is typically used for metric depth methods predicting absolute depth.

Scale alignment (s-alignment). This strategy rescales the predicted disparity using a single global scale factor. Let d denote the predicted disparity and d^* the ground-truth disparity. The scale $\tilde{s}$ is estimated as the ratio of medians, $\tilde{s} = \text{med}(d^*)/\text{med}(d)$. The aligned disparity is then $\hat{d} = \tilde{s}d$, and the corresponding depth is obtained as $\hat{z} = 1/\hat{d}$.

Scale-and-shift alignment (s&t-alignment). This alignment estimates both a scale $\tilde{s}$ and a shift $\tilde{t}$ in disparity space by solving a least-squares problem. Specifically, the optimal parameters are estimated by minimising the least-squares error, $\sum_{i=1}^{M}(sd_i + t - d_i^*)^2$, where M denotes the number of valid disparity values. The aligned disparity is then $\hat{d} = \tilde{s}d + \tilde{t}$, and the corresponding depth is obtained as $\hat{z} = 1/\hat{d}$.

We note that these alignment strategies rely on different estimation principles: scale alignment uses a robust median-based estimator, whereas scale-and-shift alignment relies on a least-squares formulation. Although this introduces a methodological inconsistency, we follow the standard evaluation protocols commonly adopted in the literature.

2.4 Perspective Projection and Local Geometric Behaviour

Perspective projection is the fundamental geometric mapping that relates a 3D scene to its 2D image under a perspective camera model. This mapping is inherently non-linear due to the division by depth. Consequently, changes in viewpoint induce geometric distortions in the image, causing the apparent shape of objects to vary, even when the underlying geometry remains unchanged. Although perspective projection itself is a mapping from 3D to 2D, the geometric distortions it induces on the image can be analysed using planar transformation models. In particular, different transformation classes preserve different geometric properties, providing a useful framework for characterising the local and global behaviour of perspective effects.

2.4.1 A Hierarchy of Transformations

Geometric transformations can be organised into a hierarchy according to the geometric properties they preserve. We introduce these transformations progressively, starting from the most restrictive class, namely Euclidean transformations (or isometries), and gradually increasing their degrees of freedom until reaching the most general class considered here, namely projective transformations. As the transformation model becomes more general, fewer geometric quantities remain invariant[1],

[1] An alternative to describing the transformation algebraically, i.e. as a matrix acting on coordinates of a point or curve, is to describe the transformation in terms of those elements or quantities that are preserved or invariant. A (scalar) invariant of a geometric configuration is a function of the configuration whose value is unchanged by a particular transformation. For example, distance

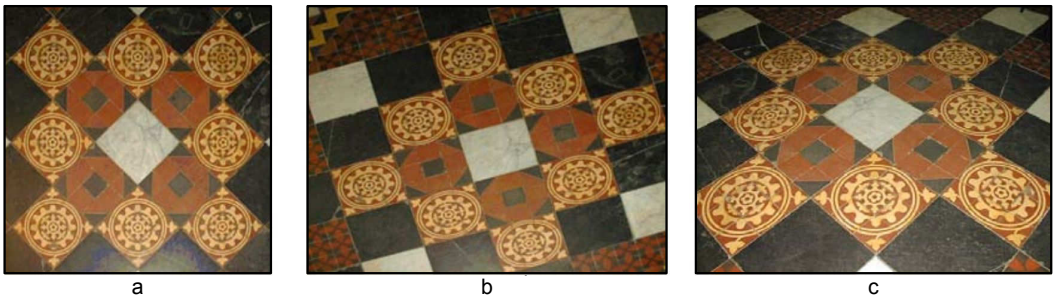


Figure 2.1: Distortions arising under central projection. Images of a tiled floor. (a) Similarity: the circular pattern is imaged as a circle. A square tile is imaged as a square. Lines which are parallel or perpendicular have the same relative orientation in the image. (b) Affine: The circle is imaged as an ellipse. Orthogonal world lines are not imaged as orthogonal lines. However, the sides of the square tiles, which are parallel in the world, are parallel in the image. (c) Projective: Parallel world lines are imaged as converging lines. Tiles closer to the camera have a larger image than those further away. Reprinted from [Hartley and Zisserman, 2003].

while the range of geometric distortions increases. The distortion effects associated with the different transformation classes are illustrated in figure 2.1, while their invariance properties are summarised in table 2.1.

Class I: Isometries. Isometries are transformations of the plane $\mathbb{R}^2$ that preserve Euclidean distance (from iso = same, metric = measure). They are composed of rotations and translations, and optionally reflections. The invariants in this class include length (the distance between two points), angle (the angle between two lines), and area.

Class II: Similarity transformations. A similarity transformation (or more simply a similarity) is an isometry composed with an isotropic scaling. A similarity transformation is also known as an equi-form transformation, because it preserves shape. Angles between lines are not affected by rotation, translation, or isotropic scaling, and are therefore similarity invariants. In particular, parallel lines are mapped to parallel lines. The distance between two points is not a similarity invariant, but the ratio of two lengths is invariant because the scaling cancels out. Similarly, the ratio of two areas is invariant because the squared scaling factor cancels out.

Class III: Affine transformations. Affine transformations generalise similarity transformations by allowing anisotropic scaling and shearing. Because an affine transformation includes non-isotropic scaling, the similarity invariants of length ratios and angles between lines are not preserved under an affinity. Three invariants in this class are: parallel lines, ratio of lengths of parallel line segments and ratio of areas.

is a Euclidean, but not a similarity, invariant. The angle between two lines is both a Euclidean and a similarity invariant.

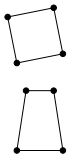
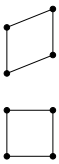
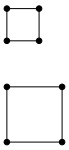
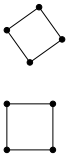
Group	Distortion	Invariant properties
Projective 8 dof		Concurrency, collinearity, cross ratio (ratio of ratio of lengths).
Affine 6 dof		Parallelism, ratio of areas, ratio of lengths on collinear or parallel lines (e.g. midpoints), The line at infinity, l_∞.
Similarity 4 dof		Ratio of lengths, angle. The circular points, I, J (see section 2.7.3).
Euclidean 3 dof		Length, area

Table 2.1: Geometric properties invariant to commonly occurring planar transformations. The distortion column shows typical effects of the transformations on a square. Transformations higher in the table can produce all the actions of the ones below. Reprinted from [Hartley and Zisserman, 2003]

Class IV: Projective transformations. Projective transformations, also called homographies, constitute the most general planar transformations considered in this hierarchy. It is a general non-singular linear transformation of homogeneous coordinates. Projective transformations preserve straight lines and incidence relations, such as collinearity and line intersections, but generally do not preserve parallelism, angles, distances, or area ratios. Their invariants include collinearity, concurrency, tangency relations, and the cross-ratio of four collinear points.

Summary and comparison. Table 2.1 summarises the 2D transformation groups and their invariant properties. Transformations lower in the table are specializations of those above. A transformation lower in the table inherits the invariants of those above. Affinities (6 Degrees of Freedom (dof)) occupy the middle ground between similarities (4 dof) and projectivities (8 dof). They generalise similarities in that angles are not preserved, so that shapes are skewed under the transformation. On the other hand their action is homogeneous over the plane: for a given affinity the

area of an object (e.g. a square) is the same anywhere on the plane. In contrast, for a given projective transformation, area scaling varies with position (e.g. under perspective a more distant square on the plane has a smaller image than one that is nearer, as shown in figure 2.1. Projective transformations preserve only straight lines but not parallelism, angles, or distances. A key property of projective transformations is that parallel lines in 3D generally converge towards vanishing points in the image. This phenomenon gives rise to perspective distortion, which is inherently viewpoint-dependent. Consequently, changes in camera pose induce spatially varying transformations of the observed scene that cannot, in general, be captured by simpler models such as affine transformations[2].

2.4.2 Local Approximation of Perspective Transformations

Although perspective projection is globally projective, it can be locally approximated by simpler transformations. At first order, the mapping can be approximated by an affine transformation, meaning that, within a sufficiently small neighbourhood, perspective projection behaves approximately affinely. Importantly, this approximation is only local: while small image neighbourhoods may behave approximately affinely, no single affine transformation can globally model a general perspective projection of a non-planar scene. In fact, this first-order approximation is insufficient to fully characterise perspective distortion. As a result, affine transformations provide only a partial description of perspective projection.

These limitations directly impact computer vision tasks such as keypoint detection and description, where the assumption of local affine invariance may break down under strong perspective effects. Many classical keypoint detectors and descriptors are designed to be invariant or covariant only to restricted transformation groups, such as similarity or affine transformations, rather than to full perspective deformations. Consequently, their robustness may degrade significantly under strong viewpoint changes on non-planar surfaces. In the context of MIS, these distortions are particularly pronounced due to close-range imaging conditions and frequent camera motion. During organ exploration, strong viewpoint variations induce significant perspective effects on organ surfaces, which can severely impact downstream computer vision tasks.

2.4.3 Conformal Transformations and Local Shape Preservation

The transformation hierarchy discussed above is mainly organised according to global invariance properties. Moving from rigid to projective transformations progressively increases the flexibility of the mapping while reducing the number of preserved geometric quantities. In this hierarchy, projective transformations are the

[2]For planar scenes, the geometric relationship between two views can be represented exactly by a homography. However, this assumption breaks down for curved surfaces, where the induced deformation varies across the image and cannot be represented globally by a single transformation.

Table 2.2: Comparison between projective transformations and conformal flattening.

Property	**Projective transformation**	**Conformal flattening**
Distance preservation	No	No
Angle preservation	No	Yes, locally
Parallelism preservation	No	Generally no
Local shape	Distorted by perspective	Preserved up to isotropic scaling
Local behaviour	Locally approximable by affinity	Locally equivalent to similarity
Typical applications	Multiview reconstruction	Texture mapping

most general global transformations considered here: they preserve straight lines, but no longer preserve distances, angles, parallelism, or local shape.

Conformal transformations occupy a different position in this discussion. Rather than being defined by global geometric invariants, they are characterised by a local angle preservation property. In this sense, they can be interpreted as transformations that behave locally like similarity transformations, even though the scaling may vary across the domain.

Conformal and projective transformations represent different properties as summarised in table 2.2. Projective transformations preserve straight lines globally, but generally distort angles, relative scale, and local shape. Conformal transformations, on the other hand, preserve angles and local shape but do not generally preserve straight lines or global structure. Consequently, while projective mappings are well suited for modelling the physical image formation process, conformal mappings provide a useful geometric framework for controlling local distortions and anisotropic deformations. For this reason, conformal parameterisations are widely used in computer graphics applications such as texture mapping and surface flattening.

As will be investigated in chapter 4, the local shape preservation property makes conformal transformations particularly relevant for mitigating perspective-induced distortions. Although they do not model perspective projection itself, they constrain the local deformation to behave as closely as possible to a local similarity transformation, thereby reducing local shape distortion.

Chapter 3

2.5D Reconstruction of the Surgical Scene from Monocular Methods or Uncalibrated Stereoscopes

Contents

3.1 Introduction and Chapter Overview

The 3D reconstruction of the surgical scene from intraoperative images is a fundamental component of vision-based surgical guidance systems. In MIS, where direct perception of the anatomy is limited, 3D reconstruction enables a geometric understanding of the observed scene and supports a range of applications, including AR [Malhotra et al., 2023, Xu et al., 2024], navigation [Grasa et al., 2013], and intraoperative measurements [Wang et al., 2018][1]. Reconstruction can be broadly categorised into full 3D reconstruction and 2.5D reconstruction. These two reconstructions differ in their representation of scene geometry. Full 3D reconstruction aims to recover a complete, viewpoint-independent geometric representation of the scene in a global reference frame. In contrast, 2.5D reconstruction estimates an intrinsically viewpoint-dependent depth map. We study 2.5D reconstruction; therefore, our objective is not to recover a full 3D representation of the scene, but rather to estimate depth maps expressed in camera reference frames. While 2.5D reconstruction occurs in multiple contexts, we consider two of the main approaches.

The first 2.5D reconstruction approach corresponds to monocular single-image reconstruction. We refer to this approach as Monocular MoSDE, which is the problem of predicting scene depth from a single image without requiring a second view, an active illumination pattern, or temporal information. MoSDE methods are flexible and widely applicable, requiring only a single RGB image and being independent of specialised hardware. Despite inherently suffering from scale ambiguity, and often from a stronger scale-and-shift ambiguity, as discussed in section 2, MoSDE remains highly attractive in MIS, as it can be readily applied to images acquired with standard monocular laparoscopes.

The second 2.5D reconstruction approach is stereoscopic reconstruction, which exploits binocular vision to estimate depth through disparity estimation. Stereo methods provide metric reconstruction under appropriate conditions, but rely on accurate camera calibration and image rectification, and are therefore sensitive to deviations from ideal acquisition settings. They are attractive because stereo cameras are being increasingly used in surgery.

In practice, neither approach is perfectly adapted to and reliable in the MIS conditions. While learning-based MoSDE methods have achieved significant progress, their performance in surgical contexts remains questionable. These methods are typically trained on non-surgical datasets (e.g., urban scenes), as ground-truth depth data in MIS is scarce. Consequently, they are strongly affected by domain shift and image perturbations present in MIS conditions such as specular reflections, smoke, and non-uniform illumination. At the same time, stereo reconstruction is challenged

[1] We note that the reconstructed geometry may be subject to scale or scale-and-shift ambiguities. Details are provided in section 2. Importantly, many downstream tasks in MIS operate effectively even in the presence of this ambiguity, including collision avoidance [Li et al., 2023], instrument tracking [Narasimhan et al., 2025], and surgical simulation [Preetha et al., 2020]. In addition, AR systems can recover the unknown parameters during the registration stage.

by imperfect calibration commonly encountered in the operating room. Accurate calibration may not be consistently available throughout surgery, as the camera parameters may vary with zoom, focus and other digital or mechanical adjustments. Moreover, calibration is also impractical to perform routinely in operating room conditions. As a result, both the intrinsic and extrinsic parameters are often unavailable or unreliable, making the standard stereo pipelines difficult to apply. Consequently, 2.5D reconstruction methods that perform well under controlled conditions may degrade significantly in real surgical environments.

This chapter addresses these challenges by investigating both approaches. First, it introduces RoDEM[2], a complete benchmark designed for a detailed accuracy and robustness analysis of state-of-the-art MoSDE methods in MIS, highlighting their limitations and failure modes. Second, the chapter presents CATS[3], a method for 2.5D reconstruction from uncalibrated stereo laparoscopes, which aims to reduce the reliance on precise camera calibration by explicitly modelling the stereo camera configuration variability during training. Taken together, this chapter addresses Research Questions RQ1 and RQ2.

3.2 The RoDEM Benchmark

This section introduces the RoDEM benchmark for evaluating monocular 2.5D reconstruction in MIS. It first reviews the main challenges of monocular depth estimation in surgical scenes, then motivates the need for a robustness-oriented benchmark. The section then describes existing datasets and benchmarks, introduces the RoDEM acquisition setup and evaluation protocol, and finally reports quantitative and qualitative results for state-of-the-art MoSDE methods.

3.2.1 Background on Monocular 2.5D Reconstruction and its Challenges in MIS

Monocular 2.5D reconstruction aims to estimate scene depth from a single RGB image. This problem has been extensively studied in computer vision for several decades. Early approaches rely on hand-crafted geometric or photometric cues that relate image appearance to scene structure under explicit physical assumptions. A representative example is Shape-from-Shading (SfS) [Horn, 1970], where surface geometry is inferred from variations in image intensity. In its classical formulation, SfS assumes Lambertian reflectance, spatially constant albedo, known lighting conditions, and sufficiently smooth surfaces. Under these assumptions, image brightness can be related to local surface normals, from which relative depth may be recovered. While elegant and physically interpretable, these assumptions are rarely satisfied in surgical environments. Endoscopic scenes exhibit complex specular reflections, and spatially varying tissue properties, which strongly limit the applicability of such

[2]Project page: https://rasoul-sharifian.github.io/RoDEM/
[3]Project page: https://rasoul-sharifian.github.io/CATS/

models. Other classical approaches exploit other cues such as defocus [Pentland, 1987], perspective effects [Szeliski, 2022], and occlusion boundaries [Zhou and Dong, 2022]. Although these methods provide valuable insight into the visual cues underlying depth perception, they are typically restricted to controlled settings and degrade when conditions deviate from idealised image formation models.

Recent progress in monocular depth estimation has been largely driven by learning-based methods, which replace explicit modelling of visual cues with data-driven approaches. Instead of relying on a limited set of assumptions, these approaches learn statistical relationships between image appearance and depth, typically from large datasets, leading to substantial improvements in robustness and accuracy, particularly in outdoor and urban environments.

Learning-based methods can be broadly categorised into supervised and self-supervised approaches. In supervised learning, models are trained using paired RGB images and ground-truth depth maps obtained from sensors such as Light Detection and Ranging (LiDAR), RGB-D cameras, structured-light systems, synthetic rendering pipelines, or multi-view reconstruction methods such as Structure-from-Motion (SfM). However, in MIS, such datasets remain scarce. Acquiring dense and reliable ground-truth depth in vivo is technically challenging, and rarely feasible at scale and at all. Furthermore, training on non-clinical or animal data introduces a significant domain gap due to differences in appearance, deformation patterns, and imaging conditions.

To overcome the scarcity of labelled data, self-supervised approaches replace explicit depth supervision with reconstruction-based objectives. Typically, depth and camera motion are jointly estimated by enforcing photometric consistency across adjacent frames or stereo views [Godard et al., 2017]. In this setting, a predicted depth map is considered accurate if it allows one view to be geometrically warped into another with minimal photometric error. This approach has proven highly effective in domains such as autonomous driving, where scenes are predominantly rigid and illumination is relatively stable. However, the assumptions underlying self-supervised approaches are fundamentally violated in MIS. In endoscopic imaging, the light source is rigidly attached to the camera and moves with it, thereby breaking the brightness constancy assumption underlying photometric losses. In addition, surgical scenes are inherently non-rigid: tissues deform continuously due to respiration, manipulation, insufflation, and surgical actions. Further challenges arise from specular reflections, smoke, bleeding, occlusions by instruments, and low-texture tissue surfaces. As a result, the assumptions of scene rigidity and photometric consistency that underpin many self-supervised approaches are frequently violated.

3.2.2 Problem Statement and Motivation for RoDEM

Despite the challenges outlined above, recent learning-based MoSDE methods have demonstrated promising performance in MIS settings [Shao et al., 2022, Cui et al., 2024, Budd and Vercauteren, 2024b]. However, these results are typically obtained

on small-scale datasets, which do not adequately capture the variability and complexity of real surgical environments. In practice, depth estimation in MIS is affected by a wide range of domain-specific factors. These include general imaging conditions, specifically, the use of a camera-embedded light source and the presence of strong specular reflections, as well as more specific intraoperative conditions, including bleeding, smoke, motion blur, and surgical interactions. These factors introduce significant distribution shifts with respect to the data used for training, and can be interpreted as *perturbations* that generate Out-of-Distribution (OoD) inputs.

Current evaluation protocols do not sufficiently account for these conditions. Existing benchmarks either rely on unrealistic data [Wang et al., 2024a], lack reliable depth annotations [Twinanda et al., 2016], or fail to cover the full set of perturbations encountered in MIS [Edwards et al., 2022, Bobrow et al., 2023]. For instance, important factors such as lens contamination, defocus blur, bleeding, or surgical actions are often omitted. Furthermore, evaluation metrics are not systematically designed to distinguish between performance under nominal conditions and robustness to perturbations [Wang et al., 2024a]. This is typically important, since high performance under nominal conditions does not guarantee stability under perturbations, which is critical for deployment in surgical environments.

As a result, there is a lack of a comprehensive and clinically relevant benchmark for assessing MoSDE methods in MIS. Such a benchmark should explicitly evaluate both (i) accuracy under clean conditions and (ii) robustness under controlled perturbations. This requires three key components: (1) a well-defined set of relevant perturbations, (2) a realistic dataset with corresponding depth annotations spanning these conditions, and (3) evaluation metrics that jointly capture accuracy and robustness.

The RoDEM benchmark is introduced to address these limitations. Concretely, RoDEM comprises three main components. First, it defines nine representative MIS perturbations, including six image perturbations—smoke, bleeding, bright and dark illumination variations, defocus blur, motion blur, and lens dirtiness—and two surgical gesture perturbations, namely organ deformation and incision. Second, to evaluate the generalisability of the methods, it is based on collecting a new ex-vivo dataset on which the methods are evaluated. This dataset comprises RGB laparoscopic images of sheep organs in three settings (liver, kidney, and heart–lung), with per-image ground-truth depth acquired using a structured-light camera. It includes 44 video sequences (1,156 seconds in total) and 21 image sets with 906 frames, amounting to 29,803 annotated endoscopic images covering all considered perturbations. Third, it employs rigorously defined evaluation metrics to jointly assess accuracy and robustness, and benchmarks nine existing learning-based MoSDE methods. The design of RoDEM and its evaluation protocol are given in the following sections.

3.2.3 Review of Existing Datasets

In this section, we review four representative datasets providing depth information for MIS. While a larger number of surgical image datasets exist (e.g., [Modrzejewski et al., 2019]), they are not considered here as they do not provide depth maps and are therefore not suitable for quantitative evaluation of 2.5D reconstruction methods.

SCARED. The SCARED dataset [Allan et al., 2021a][4] (EndoVis challenge, MICCAI 2019) is a stereo endoscopic dataset acquired on porcine cadavers. It relies on structured-light projection at selected keyframes to enable dense depth reconstruction. During acquisition, the endoscope is manipulated using a da Vinci Xi robot, and camera poses are recorded through the robot's internal sensors. Depth for intermediate frames is obtained by transferring and reprojecting 3D points from keyframes using these poses.

This process assumes a static scene between keyframes, which prevents capturing organ deformation or surgical actions such as incisions. Moreover, the dataset does not include typical MIS perturbations. In addition, as the camera moves away from keyframes, the resulting depth maps become increasingly sparse, and accumulated pose inaccuracies lead to noticeable misalignment between images and depth ground-truth.

SERV-CT. The SERV-CT dataset [Edwards et al., 2022] is a stereo endoscopic dataset in which depth ground-truth is obtained from cone-beam CT. Acquisition is performed by placing porcine cadavers inside a CT scanner, ensuring that both the endoscope and the scene are visible in the scan. Depth is generated by manually aligning the CT-based 3D reconstruction with stereo images.

This acquisition pipeline is complex, as it requires synchronised CT scans and stereo image pairs, along with manual alignment. As a result, the dataset is limited to only 16 image–depth pairs. In addition, it is acquired under controlled conditions and does not include realistic MIS perturbations.

Hamlyn. The Hamlyn dataset [Mountney et al., 2010, Stoyanov et al., 2005, 2010, Pratt et al., 2010] is a laparoscopic video dataset containing both in-vivo and ex-vivo sequences, with challenging conditions such as weak texture, deformation, specularities, occlusions, and surgical tools. To make the dataset suitable for depth estimation, [Recasens et al., 2021] generated ground-truth depth maps using the Libelas stereo matching software [GeigerandAndreas et al., 2010].

However, Libelas uses sparse estimated depth from keypoint matches to interpolate a dense depth map, making it inaccurate due to missing details and smoothing sharp depth transitions, and leaving blank areas at the image borders and within the image.

EndoAbs. The EndoAbs dataset [Penza et al., 2018] is a stereo endoscopic dataset captured using phantom abdominal organs. The depth ground-truth, aligned with

[4] https://endovissub2019-scared.grand-challenge.org/

the left camera, is generated using a laser scanner. The dataset was captured under different conditions, namely 3 different light levels, smoke, and two phantom-endoscope distances (5 cm and 10 cm).

Despite these controlled variations, the dataset remains limited in size, comprising only 120 image–depth pairs. More importantly, phantom-based environments cannot fully reproduce the visual complexity and variability of real surgical scenes.

3.2.4 Comparison and Limitations of Existing Datasets

The main characteristics of existing datasets are summarised in table 3.1. A key limitation is the small number of image–depth pairs available in most datasets, which raises concerns regarding their suitability for robust benchmarking.

Datasets relying on CT or laser scanning (i.e., SERV-CT, EndoAbs) provide accurate depth ground-truth but impose strong acquisition constraints. These systems require several seconds to minutes per acquisition, during which the scene must remain static. As a result, they cannot capture dynamic surgical scenes or incorporate image-based perturbations, thereby limiting dataset scalability.

In contrast, stereo-based approaches (i.e., Hamlyn) enable large-scale data collection and video sequences, but at the cost of reduced depth accuracy, particularly in low-texture regions. Structured-light methods (i.e., SCARED) improve reconstruction in such regions and offer a compromise between accuracy and scalability. However, they require projecting patterns onto the scene, which alters its appearance and prevents continuous acquisition. Additionally, keyframe-based depth transfer still relies on static scene assumptions and interrupts the acquisition process.

Another fundamental limitation across all existing datasets is the absence of realistic surgical perturbations. Most datasets are acquired under controlled conditions, with stable illumination and limited scene dynamics. As a result, they fail to capture the degradations commonly encountered in real surgical environments.

3.2.5 Review of Existing Benchmarks

Only a limited number of studies have investigated the robustness of MoSDE methods. In the context of urban scenes, prior work has introduced synthetic image corruptions to evaluate robustness by applying predefined perturbations to clean images [Kong et al., 2024]. However, these studies are restricted to non-medical datasets and rely exclusively on artificially generated degradations. A similar approach [Wang et al., 2024a] has been explored in the context of MIS, where synthetic perturbations are applied to endoscopic images to benchmark MoSDE methods. While this represents an initial step toward robustness evaluation in surgical settings, the use of synthetic corruptions remains a major limitation. Such perturbations do not accurately reproduce the complex visual phenomena observed in real surgery, including dynamic smoke, fluid interactions, and specular reflections. Furthermore, existing benchmarking efforts remain limited in scope, typically eval-

Table 3.1: Comparison of existing MIS datasets including depth ground-truth. The comparison shows the presence of image perturbations (S: Smoke, M: Motion blur, L: Light variation, D: Lens dirtiness, F: Defocus blur, B: Bleeding) and sequences (C: Camera movement, D: Deformation, I: Incision).

Dataset	Data type	Zone	Img-depth pairs	Depth acquisition	Image perturbations						Sequences		
					S	M	L	D	F	B	C	D	I
SCARED[5]	Ex-Vivo	Abdomen	45	Structured light	-	-	-	-	-	-	✓	-	-
SERV-CT	Ex-Vivo	Abdomen, Thorax, Pelvis	16	CT scan	-	-	-	-	-	-	-	-	-
Hamlyn	Ex-Vivo, In-Vivo, Phantom	Abdomen, Pelvis	92694	Stereo matching	-	-	-	-	-	-	✓	-	-
EndoAbs	Phantom	Abdomen	120	Laser scanner	✓	-	✓	-	-	-	-	-	-
RoDEM (proposed)	Ex-Vivo	Thorax, Abdomen	29803	Depth sensor	✓	✓	✓	✓	✓	✓	✓	✓	✓

uating only a small number of methods, thus ignoring recent advances, including foundation models and their fine-tuned variants.

3.2.6 RoDEM Dataset Acquisition Setup

The proposed RoDEM dataset[6] addresses the limitation of previous works through a different acquisition strategy. Depth ground-truth is obtained using a dedicated depth sensor capable of providing dense measurements at full frame rate. These depth maps are automatically aligned with laparoscopic images through a calibrated transformation. This approach enables acquisition under realistic conditions without interrupting the recording process and without relying on synthetic simulation. It allows capturing both single images and continuous sequences, including dynamic events such as organ deformation and surgical incisions. Although depth sensors may provide lower absolute accuracy than CT or laser-based systems, they offer a practical trade-off between accuracy and scalability, enabling large-scale data collection in realistic settings.

The acquisition setup is illustrated in figure 3.1. The main components consist of an endoscope (Column: Storz Image 1 connect & Storz Image 1 H3 link; Camera: Storz Image 1 HD (H3-Z TH100); Light source: Storz SCB xenon 175 - 20132120) and an RGB-D camera (Intel RealSense D435i). While the RGB-D camera provides depth measurements, it cannot reproduce the optical characteristics of surgical endoscopes, including viewpoint, illumination, and image appearance. Conversely, the laparoscope provides clinically realistic images but no direct depth information.

[6]RoDEM dataset: https://drive.google.com/drive/folders/1d3e5y7ijcob1_qhPa6p1gLYR5bcJMdPj

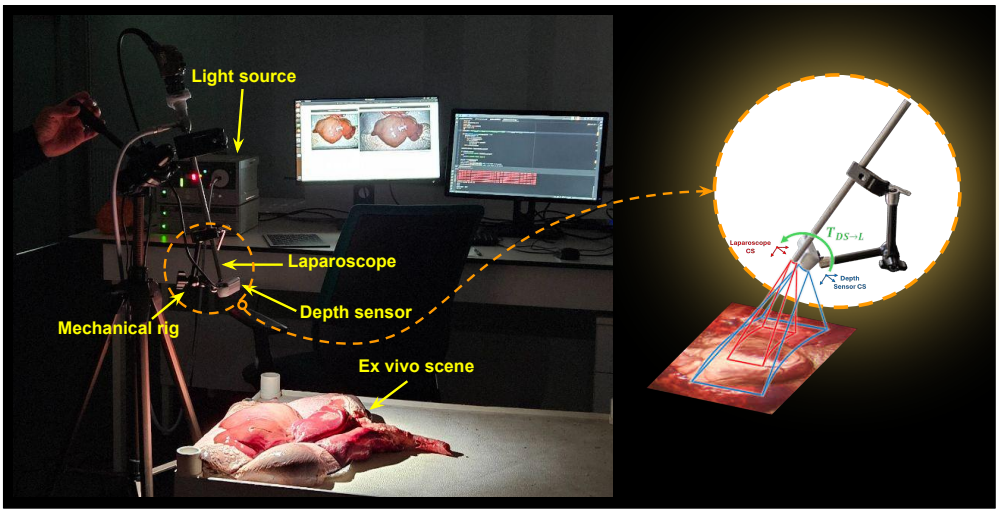


Figure 3.1: Left, the proposed RoDEM dataset acquisition setup. Right, schematic of the depth sensor, laparoscope, their fields of view, coordinate systems, and the rigid transformation between them.

Combining both devices therefore enables the acquisition of accurate depth maps together with visually representative laparoscopic images. The two cameras are rigidly attached using a mechanical rig to ensure a fixed transformation between their coordinate systems. The acquisition process is structured into three phases.

1) Pre-acquisition phase: Prior to data acquisition, both cameras were calibrated and the rigid transformation between them was estimated. Synchronously, images of a ChArUco board were captured using both the endoscope and the RGB-D camera. OpenCV was used to perform intrinsic calibration and to estimate the rigid transformation between the depth sensor coordinate system (DS) and the laparoscope coordinate system (L), denoted as $T_{DS \to L}$. Let P_L denote a 3D point in the laparoscope coordinate system and P_{DS} the same 3D point in the depth sensor coordinate system. The transformation between the two systems is estimated from the ideal constraint $P_L = T_{DS \to L}(P_{DS})$, by minimising the least-squares reprojection transfer error. Calibration accuracy was evaluated using the reprojection error. This was computed by back-projecting the ChArUco IDs observed in the RGB-D image into 3D space, followed by reprojection into the laparoscope image using the estimated transformation. An average reprojection error of 1.21 pixels was obtained over 313 calibration images of resolution 640×480. The depth accuracy of the system was further evaluated using a 3D-printed liver phantom with known geometry. The average error between the phantom Computer-Aided Design (CAD) model and the reconstructed 3D model was measured to be 1.4 mm.

2) Acquisition phase: During the acquisition phase, scenes were recorded simultaneously using both cameras. Depth maps were obtained from the RGB-D camera, while laparoscopic images were recorded at a resolution of 640×480 pixels. Data were collected across three anatomical settings: liver, kidney, and heart–lung, using ex-vivo organs from a single sheep. Each setting was subjected to real image perturbations, generated as follows:

Smoke: generated using an electrosurgical device to produce realistic surgical smoke within the scene.

Bleeding: simulated by applying artificial blood spray to the organ surface.

Bright and dark illumination variations: obtained by adjusting the intensity of the laparoscopic light source.

Defocus blur: induced by modifying the focus settings of the laparoscope.

Motion blur: caused by rapid camera motion during acquisition.

Lens dirtiness: simulated by spraying the laparoscope lens.

In addition, two surgical gesture perturbations were introduced:

Organ deformation: induced by applying forces to the organs using surgical tools.

Incision: performed by making cuts in the tissue using standard surgical instruments.

3) Post-acquisition phase: In the post-acquisition phase, depth maps were aligned with the laparoscopic images, and each image was assigned a perturbation severity level. Concretely, using the transformation estimated during calibration, depth data from the RGB-D camera were mapped into the laparoscope coordinate system. Z-buffering was applied to remove depth values corresponding to occluded regions in the laparoscopic view. Surgical tools were segmented using Medical SAM 2 [Zhu et al., 2024] and removed from the depth maps to mitigate inaccuracies caused by depth sensor limitations in capturing the depth of surgical tools, particularly in regions close to the camera. Additionally, depth values were clamped to a range between 1 cm and 50 cm, reflecting typical MIS operating conditions. Images were categorised into different severity levels based on visual inspection. Representative examples of the perturbations are shown in figure 3.2. Additional samples, including laparoscopic images with corresponding depth maps and reconstructed point clouds, are shown in figure3.3.

3.2.7 Benchmarking

Benchmarked methods. We benchmark a representative set of recent MoSDE methods using the RoDEM dataset, including MiDaS V3.1 [Birkl et al., 2023], Depth Anything V1 (DA) [Yang et al., 2024a], Depth Anything V2-Large (DA2) [Yang

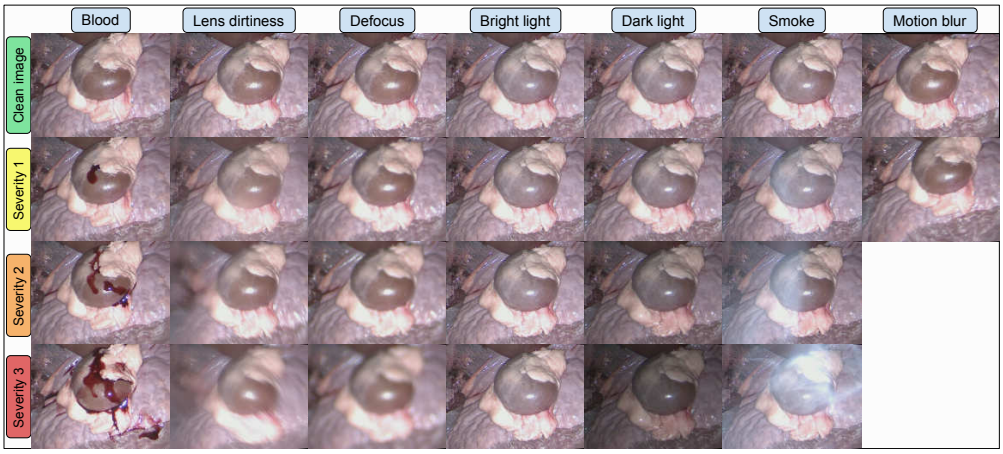


Figure 3.2: Representative images illustrating the identified perturbations and their respective severity levels, selected from a total of 9,035 images in the kidney setting of the RoDEM dataset.

Table 3.2: Detailed information on the RoDEM dataset scenarios and perturbations. 'Pairs' refer to the number of image and depth ground-truth pairs, 'Sev.' to the number of severity levels and 'Seq.' to the number of video sequences.

	Scene 1			Scene 2			Scene 3		
	Pairs	Sev.	Seq.	Pairs	Sev.	Seq.	Pairs	Sev.	Seq.
Blood	19	3	-	65	3	-	46	3	-
Lens Dirtiness	15	3	-	12	3	-	27	3	-
Defocus Blur	32	3	-	37	3	-	30	3	-
Light (Bright)	12	3	-	12	3	-	12	3	-
Light (Dark)	11	3	-	11	3	-	11	3	-
Smoke	282	3	-	123	3	-	77	3	-
Motion Blur	22	1	-	25	1	-	25	1	-
Deformation	5575	-	12	3450	-	7	1178	-	4
Incision	5322	-	7	5300	-	5	8072	-	9
Subtotal	11290	-	19	9035	-	12	9478	-	13
Total Pairs	29803								
Total Sequences	44								

et al., 2024b], AF-SfM learner [Shao et al., 2022], EndoDAC [Cui et al., 2024], Transferring Relative Monocular Depth to Surgical Vision (TRMDSV) [Budd and Vercauteren, 2024b], ZoeDepth [Bhat et al., 2023], Depth Anything V2-metric (DA2-metric) [Yang et al., 2024b], and UniDepth [Piccinelli et al., 2024].

MiDaS, DA, and DA2 correspond to state-of-the-art foundation models trained on large-scale, non-MIS datasets. These models predict disparity rather than absolute depth and are trained using scale- and shift-invariant losses. As a result, their outputs represent inverse depth up to an unknown scale and shift.

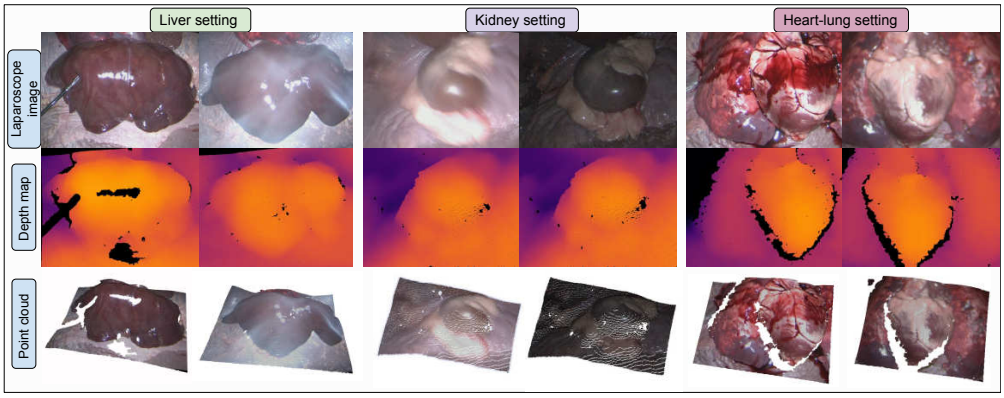


Figure 3.3: Samples of laparoscope images, along with their corresponding depth maps and point clouds, for the three settings used in the RoDEM dataset.

AF-SfM learner, EndoDAC, and TRMDSV are specifically designed for MIS images. AF-SfM learner is trained on the SCARED dataset and explicitly accounts for brightness inconsistencies by estimating an appearance flow. EndoDAC and TRMDSV are both fine-tuned versions of Depth Anything, but follow different strategies. EndoDAC is trained on SCARED and incorporates an adaptation mechanism that estimates camera intrinsics, enabling training from uncalibrated images. In contrast, TRMDSV leverages a broader range of medical datasets and employs temporal consistency-based self-supervision to improve generalisation.

ZoeDepth, DA2-metric, and UniDepth are metric MoSDE models that directly predict absolute depth. In principle, these models do not require additional alignment between predictions and ground-truth during evaluation.

Overall, this selection provides a comprehensive evaluation set, covering general-purpose foundation models, MIS-specific approaches, and metric depth estimation methods.

Prediction-to-Ground-Truth alignment. A key consideration when evaluating MoSDE methods is the alignment between the predicted outputs and the ground-truth depth. Several methods predict disparity rather than absolute depth, resulting in outputs defined up to an unknown scale and shift. Consequently, accurate evaluation requires estimating the optimal scale and shift parameters to align the predicted disparity with the ground truth prior to computing depth metrics. Several alignment strategies have been proposed in the literature [Birkl et al., 2023]. We categorise them into three groups: no alignment, scale alignment (s-alignment), and scale-and-shift alignment (s&t-alignment). More details are provided in section 2.3.3. After alignment, the predicted depth values are clamped to the range [1 cm, 50 cm] to reflect typical MIS operating conditions.

Metrics. We first review the depth metrics, followed by the introduction of two complementary metrics that capture the robustness of depth estimation to perturbations.

Depth metrics. After aligning the predictions with the ground-truth, depth error and depth accuracy metrics are computed. Commonly used metrics for depth evaluation include Absolute Relative Error (Abs Rel), Squared Relative Error (Sq Rel), Root Mean Squared Error (RMSE), logarithmic RMSE, and Scale-Invariant Logarithmic Error (SILog), together with the accuracy measure δ_t. In our experiments, we evaluated all of these metrics and observed a strong correlation between them, with no change in the ranking of the benchmarked methods. For this reason, we retained Abs Rel and δ_t as the representative error and accuracy metrics:

$$\text{Abs Rel} = \frac{1}{M}\sum_{i=1}^{M}\frac{|z_i^* - \hat{z}_i|}{z_i^*} \tag{3.1}$$

$$\delta_t = \frac{1}{M}\left|\left\{\hat{z}_i; i = 1, 2, \ldots, M \mid \max\left(\frac{z_i^*}{\hat{z}_i}, \frac{\hat{z}_i}{z_i^*}\right) < 1.25^t\right\}\right|, \tag{3.2}$$

where $\hat{z}_i$ and z_i^* are the aligned-predicted and ground-truth depth maps respectively. Conventionally, the value of t, which determines the 'resolution' at which accuracy δ_t is measured, is set to $t \in \{1, 2, 3\}$. In our experiments, these values for t resulted in all accuracy values being very close to 1. We thus extended the range of t to $\{0.25, 0.5, 1, 2, 3\}$ and, after careful inspection, chose $t = 0.25$ for the evaluations.

Robustness. We evaluate robustness using perturbation error and resilience rate. Perturbation error measures the degradation of a given method relative to a baseline model under perturbed conditions. Following prior robustness benchmarks [Kong et al., 2024, Hendrycks and Dietterich, 2019], we used the Mean Corruption Error (mCE) as the primary comparative metric. Concretely, Corruption Error (CE) is first computed across severity levels for a given perturbation and scene, then averaged across all perturbation types and all scenes to obtain mCE.

$$\text{CE}_{i,j} = \frac{\sum_{l=1}^{L} E_{i,j,l}}{\sum_{l=1}^{L} E_{i,j,l}^{\text{baseline}}}, \quad \text{mCE} = \frac{1}{CN}\sum_{i=1}^{C}\sum_{j=1}^{N}\text{CE}_{i,j}. \tag{3.3}$$

To normalise the effect of perturbation severity, AF-SfM learner was selected as the baseline model. In our experiments, the number of scenes was three, the number of perturbation types was nine, and the number of severity levels was three, except for motion blur, for which only one severity level was considered.

In addition to mCE, we used the Mean Resilience Rate (mRR) as a relative robustness measure [Kong et al., 2024]. This metric quantifies how much of the clean-condition performance is retained under perturbation.

Table 3.3: Accuracy and robustness, no-alignment.

	Abs Rel			$1-\delta_{0.25}$		
	Clean Error ↓	mCE ↓	mRR ↑	Clean Error ↓	mCE ↓	mRR ↑
AF-SfM	**0.852**	**1.000**	0.980	1.000	1.000	-
EndoDAC	<u>0.884</u>	<u>1.035</u>	0.991	1.000	1.000	-
TRMDSV	1.020	1.200	**1.161**	1.000	1.000	1.120
DA	0.999	1.170	1.075	1.000	1.000	-
DA2	1.000	1.171	<u>1.105</u>	1.000	1.000	-
MiDaS	1.000	1.171	1.071	1.000	1.000	-
DA2M	3.169	3.534	0.913	0.999	1.000	-
ZoeDepth	6.139	7.291	1.006	1.000	1.000	-
UniDepth	4.323	5.154	1.010	1.000	1.000	-

Table 3.4: Accuracy and robustness, s-alignment.

	Abs Rel			$1-\delta_{0.25}$		
	Clean Error ↓	mCE ↓	mRR ↑	Clean Error ↓	mCE ↓	mRR ↑
AF-SfM	<u>0.062</u>	1.000	0.984	0.460	1.000	0.880
EndoDAC	**0.050**	**0.713**	0.998	**0.358**	**0.720**	0.991
TRMDSV	0.374	5.184	0.980	0.836	1.670	0.984
DA	0.236	3.181	1.002	0.765	1.483	<u>1.101</u>
DA2	0.297	3.983	<u>1.004</u>	0.842	1.658	1.092
MiDaS	0.329	4.638	0.979	0.830	1.675	0.953
DA2M	0.068	<u>0.927</u>	1.002	<u>0.451</u>	<u>0.909</u>	1.003
ZoeDepth	0.113	1.433	**1.015**	0.685	1.254	**1.224**
UniDepth	0.089	1.284	0.998	0.591	1.198	1.000

$$\mathrm{RR}_{i,j} = \frac{\sum_{l=1}^{L} 1 - E_{i,j,l}}{L(1 - E_{\mathrm{Clean}})}, \quad \mathrm{mRR} = \frac{1}{CN} \sum_{i=1}^{C} \sum_{j=1}^{N} \mathrm{RR}_{i,j}, \tag{3.4}$$

Here, the clean error corresponds to the error measured on unperturbed data. For deformation and incision, the scene before the perturbation is treated as the clean reference, while the subsequent frames are considered perturbed.

Experimental results. Tables 3.3, 3.4 and 3.5 report clean error, mCE, and mRR for the three alignment strategies, using both Abs Rel and $1-\delta_{0.25}$. Several observations can be made regarding clean error and overall accuracy. First, the choice of alignment strategy has a substantial impact on the results. This is expected, as the benchmarked methods are trained under different assumptions: some predict metric depth directly, whereas others predict disparity up to an unknown scale or up to both scale and shift. Second, all methods encounter difficulties in producing

Table 3.5: Accuracy and robustness, s&t-alignment.

	Abs Rel			$1 - \delta_{0.25}$		
	Clean Error ↓	mCE ↓	mRR ↑	Clean Error ↓	mCE ↓	mRR ↑
AF-SfM	0.043	1.000	0.993	0.297	1.000	0.902
EndoDAC	0.033	0.753	**0.997**	0.176	0.622	0.957
TRMDSV	**0.021**	**0.499**	**0.997**	**0.039**	**0.195**	**0.971**
DA	0.024	0.564	**0.997**	0.069	0.319	0.958
DA2	0.032	0.711	**0.997**	0.172	0.559	0.968
MiDaS	0.033	0.736	**0.997**	0.185	0.593	0.961
DA2M	0.032	0.747	0.995	0.164	0.633	0.935
ZoeDepth	0.035	0.778	**0.997**	0.199	0.644	0.957
UniDepth	0.036	0.834	0.996	0.203	0.712	0.942

accurate metric depth without alignment, including those explicitly presented as metric foundation models. This suggests that metric models still require fine-tuning on MIS data in order to achieve reliable metric depth estimation in this domain.

Third, under scale alignment, EndoDAC achieves the best overall performance. This result is consistent with its design, as it is based on the Depth Anything backbone, is fine-tuned on MIS data, and is trained with a scale-alignment formulation. Fourth, under scale-and-shift alignment, TRMDSV obtains the best results, closely followed by DA. This may be explained by the fact that TRMDSV is trained with a scale-and-shift alignment strategy and benefits from exposure to a more diverse set of medical datasets, which supports stronger zero-shot transfer. Notably, DA, despite not being trained on surgical images, also demonstrates strong generalisation to the RoDEM dataset. Since these methods are trained with scale- and shift-invariant losses, their lower performance under scale-only alignment is expected.

With respect to robustness, mCE and mRR provide complementary information. The mCE metric is intended for direct comparison of methods under perturbation. Under scale alignment, EndoDAC yields the lowest mCE, whereas under scale-and-shift alignment, TRMDSV achieves the lowest value. By contrast, mRR measures the ratio of accuracy degradation from a clean condition to the perturbed one. It should therefore be interpreted with care: a high mRR does not necessarily indicate good absolute performance, as a method may simply perform poorly in both clean and perturbed settings. Taking this into account, the methods that combine strong clean accuracy with favourable mRR are again EndoDAC under scale alignment and TRMDSV under scale-and-shift alignment. These two methods therefore emerge as the most robust overall.

Figure 3.4 presents radar charts illustrating robustness across perturbation types, using the scale-and-shift alignment setting. These results show that TRMDSV and DA perform best for most perturbations. Across nearly all methods, lens dirtiness and defocus blur appear to be the most challenging perturbations, whereas motion

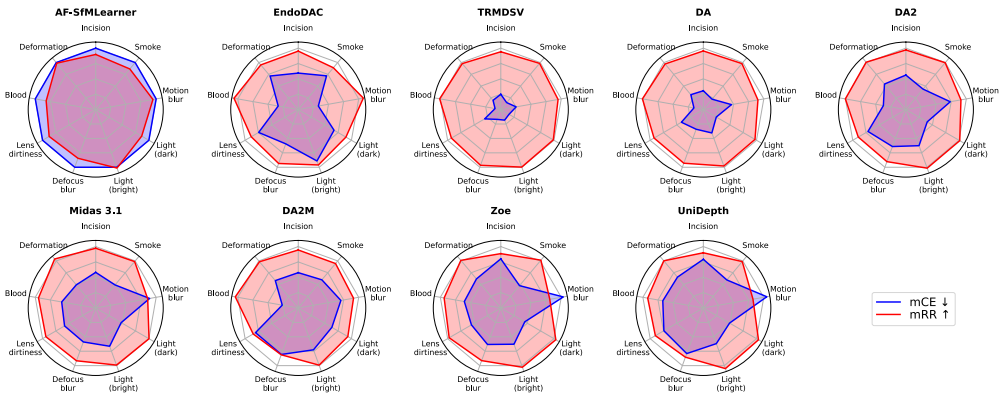


Figure 3.4: Radar charts with perturbation error and resilience rate for the nine benchmarked methods.

blur and bright light variation are the least detrimental. Additional analyses are presented in the appendix C.

3.2.8 Discussion

The proposed RoDEM dataset addresses key limitations of existing datasets and benchmarks through a fundamentally different acquisition strategy. In contrast to prior work relying on synthetic data or sparse ground-truth, RoDEM leverages a dedicated depth sensing system capable of providing dense depth measurements at full frame rate. These depth maps are automatically aligned with laparoscopic images via a calibrated transformation, enabling continuous and consistent acquisition. This design allows data collection under realistic surgical conditions without interrupting the workflow or relying on simulation assumptions. As a result, both single images and temporally coherent sequences are captured, including dynamic events such as organ deformation, surgical manipulation, and incision.

A central advantage of this approach is its ability to bridge the gap between realism and scalability. While alternative modalities such as CT or laser scanning may provide higher absolute accuracy, they are impractical for large-scale, temporally dense acquisition in surgical settings. Depth sensors, despite their lower precision, offer a favourable trade-off by enabling efficient data collection across a wide range of scenarios. This makes it possible to construct a dataset of substantial size and diversity, across multiple ex-vivo configurations. Importantly, the dataset includes a broad spectrum of clinically relevant perturbations, capturing the variability encountered in real surgical practice.

An important strength of this benchmark is that RoDEM is used exclusively for evaluation, and not for training or validation, thereby avoiding overfitting and enabling a more reliable assessment of accuracy and robustness. The dataset was deliberately designed to minimise evaluation bias as much as possible. The dataset

is publicly released, allowing the community to use it not only for benchmarking but also for training in future work.

Despite these strengths, several limitations should be acknowledged. First, the use of ex-vivo data introduces an inherent domain gap with respect to in-vivo surgical conditions. Although care was taken to reduce this gap—through the use of fresh animal organs and the introduction of realistic perturbations—the absence of physiological processes such as respiration and more complex tissue interactions remains a limiting factor. Second, while the proposed depth sensing system provides dense measurements, it is inherently subject to sensor-specific limitations, including noise, missing values, and reduced accuracy under challenging visual conditions. These imperfections may propagate to the ground-truth and affect the evaluation.

Overall, RoDEM provides a foundation for realistic and systematic evaluation in MIS, and its public availability supports future developments in both benchmarking and model improvement.

3.3 Reconstruction from Uncalibrated Stereo in MIS

This section addresses 2.5D reconstruction from uncalibrated stereo in MIS. It first reviews the principles of stereo reconstruction and the challenges arising from imperfect calibration in surgical settings. It introduces camera augmentation, analyses stereo laparoscope design and perception constraints, and presents the CATS framework for uncalibrated stereo. Finally, it describes the experimental setup and reports quantitative and qualitative results.

3.3.1 Background on Stereo 2.5D Reconstruction and its Challenges in MIS

The RoDEM benchmark demonstrates that 2.5D reconstruction can be achieved, to a certain extent, from monocular laparoscopic images. However, stereo depth estimation offers a more reliable alternative, as it provides stronger geometric constraints through triangulation based on left–right correspondences, even if not explicitly computed by all methods. This reduces the inaccuracies commonly occurring in monocular depth estimation where the models heavily rely on learnt priors. In addition, the use of stereo laparoscopes has been steadily increasing, owing to the rise of robot-assisted MIS systems.

As explained in chapter 2, a stereo system comprises two cameras that observe the scene from slightly different viewpoints, enabling depth perception through triangulation of corresponding image points. To exploit this geometric relationship, the system must first be calibrated. The stereo calibration process determines the intrinsic parameters of each camera, as well as the extrinsic parameters describing their relative rotation and translation. As illustrated in figure 3.5, in the calibrated setting, stereo methods follow a standard preprocessing pipeline including image

undistortion, rectification, and principal point alignment. Within this rectified geometry, disparity can be efficiently estimated and subsequently converted into metric depth using the calibration parameters.

This framework has been extensively studied, with a focus on improving disparity estimation accuracy and robustness. While early works established the fundamental components [Scharstein and Szeliski, 2002], recent neural architectures such as RAFTStereo [Lipson et al., 2021], IGEV++ [Xu et al., 2025] and FoundationStereo [Wen et al., 2025] have shown substantially improved performance and, importantly, generalisation across datasets. These models require rectified images with centred principal points, which necessitate camera calibration; as expected, their performance directly depends on the accuracy of calibration parameters. Calibrated stereo methods specifically adapted to MIS were proposed, including MS-DESIS [Psychogyios et al., 2022] and MCF-SMSIS [Wu et al., 2024] which jointly estimate disparity and segmentation with shared features. Beyond the medical domain, recent stereo matching methods have demonstrated promising cross-domain generalisation [Lipson et al., 2021, Wen et al., 2025, Xu et al., 2025]. However, their applicability to MIS remains constrained by domain-specific challenges, including complex illumination, tissue deformation, and, importantly, the limited availability or reliability of camera calibration in practical settings.

In practice, accurate calibration may not be available throughout surgery, as the parameters may vary with zoom level, focusing and any other digital or mechanical adjustments. Calibration is also impractical to perform routinely in operating room conditions. Consequently, both the intrinsic and extrinsic parameters are often unavailable or unreliable, making the standard stereo pipeline inapplicable.

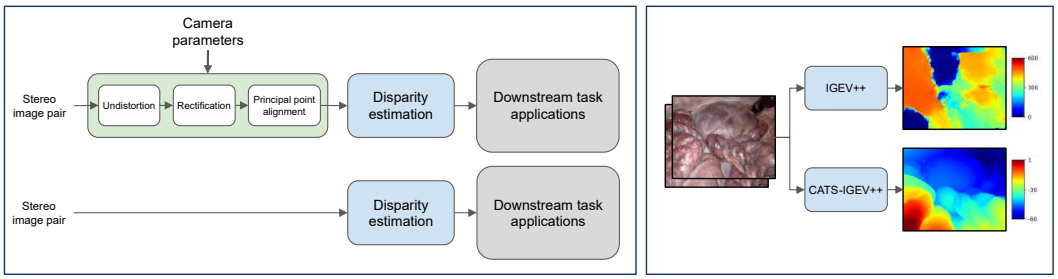


Figure 3.5: Left: Pipelines for calibrated (top) and uncalibrated (bottom) stereo matching. Right: disparity estimated by the IGEV++ [Xu et al., 2025] and the proposed CATS-IGEV++ models in uncalibrated setting.

3.3.2 Problem Statement and Motivation for CATS

Despite these well-known practical calibration challenges, uncalibrated stereo depth estimation remains largely unexplored in MIS. As we show in this section, although differences in scene content may not pose a major issue for large learning-based models, discrepancies in camera settings lead to systematic failures. This limitation

cannot be addressed by simply fine-tuning models on limited depth-labelled MIS datasets.

In the uncalibrated setting (figure 3.5), disparity can no longer be mapped to metric depth, as this relationship depends on calibration parameters. However, it still provides relative depth information. Importantly, many downstream tasks in MIS operate effectively using relative depth rather than metric depth, including collision avoidance [Li et al., 2023], instrument tracking [Narasimhan et al., 2025], and surgical simulation [Preetha et al., 2020]. Furthermore, augmented reality systems can recover scale during registration, highlighting the practical usefulness of relative depth.

To address the problem of uncalibrated stereo in MIS, we seek to leverage large-scale non-medical synthetic stereo datasets. A major limitation of these datasets is that they are generated under ideal conditions, with perfectly rectified images and centred principal points. Moreover, they are typically released without full 3D scene information, preventing re-rendering. To overcome these limitations, we introduce CATS, a framework for robust stereo learning in MIS under uncalibrated conditions. The contributions are threefold. First, we introduce camera augmentation, which generates new training samples by perturbing camera orientation and intrinsic parameters, without requiring access to scene geometry or depth. Second, we use this strategy to transform idealised stereo datasets into data representative of uncalibrated, yet near-rectified, stereo laparoscope configurations, based on a statistical modelling of real systems. Third, we demonstrate that CATS enables standard stereo architectures, such as RAFTStereo and IGEV++, to be adapted to uncalibrated settings, achieving robust 2.5D reconstruction in MIS without task-specific retraining on limited medical data. The following sections describe the proposed framework and its evaluation on surgical datasets.

3.3.3 Camera Augmentation

We introduce *Camera Augmentation* as a principled extension of conventional image augmentation, where new samples are generated by perturbing camera parameters rather than applying purely image-plane transformations. In contrast to standard augmentations, such as rotations, translations, or scalings performed directly in the image domain, Camera Augmentation explicitly models the geometric image formation process. This enables the generation of novel, geometrically consistent image samples, thereby enriching datasets that lack sufficient variability in camera orientation, focal length, principal point, or distortion.

Generating new images in a geometrically accurate manner is generally a challenging task, as it typically requires full rendering of the scene, involving knowledge of lighting conditions, surface geometry, reflectance properties, and camera parameters. This complex process is incompatible with the typical practical conditions in which most of these parameters are unavailable. In camera augmentation, we bypass this complex procedure by simply keeping the original camera position un-

changed, which nonetheless leaves a variety of other geometric controls, subsuming conventional augmentation. These new controls allow the generation of images that preserve 3D-geometric consistency without requiring complex parameters and computations, yet still improve model robustness to changes in camera geometry.

As discussed in chapter 2, We model the original camera using its intrinsic parameters $\mathbf{K} \in \mathbb{R}^{3\times3}$ and its distortion parameters $\boldsymbol{\theta} \in \mathbb{R}^k$. Choosing as 3D coordinate frame $\mathcal{C}$ the standard coordinate frame related to this camera (origin centred on the projection centre, Z axis coincident with the optical axis, and X, Y axes parallel to the retinal plane axes), the camera pose is the identity rotation $\mathbf{I}$ and zero translation $\mathbf{0}$. We model the augmented image camera using its intrinsics $\mathbf{K}' \in \mathbb{R}^{3\times3}$, its distortion parameters $\boldsymbol{\theta}' \in \mathbb{R}^k$ and its pose in $\mathcal{C}$ as the rotation $\mathbf{R}$. We leave, as already stated, the translation to $\mathbf{0}$ in order to keep the camera position unchanged.

Without distortion, for $\mathbf{P}$ the Euclidean coordinates of a 3D point, and $\tilde{\mathbf{P}}$ its homogeneous coordinates, the projections are $\tilde{\mathbf{p}} \sim \mathbf{K}\begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix}\tilde{\mathbf{P}} \sim \mathbf{K}\mathbf{P}$ and $\tilde{\mathbf{p}}' \sim \mathbf{K}'\begin{bmatrix}\mathbf{R} & \mathbf{0}\end{bmatrix}\tilde{\mathbf{P}} \sim \mathbf{K}'\mathbf{R}\mathbf{P}$, where '$\sim$' means equality up to scale. By inverting the first projection and substituting into the second one, we obtain the well-known homographic relationship $\tilde{\mathbf{p}}' \sim \mathbf{H}\tilde{\mathbf{p}}$, with $\mathbf{H} \sim \mathbf{K}'\mathbf{R}\mathbf{K}^{-1}$. We can factor the homography as $\mathbf{H} = \mathbf{K}_{\text{rel}}\mathbf{H}_{\text{R}}$ where $\mathbf{K}_{\text{rel}} = \mathbf{K}'\mathbf{K}^{-1}$ and $\mathbf{R}$ represent the sought *relative intrinsic* and *relative rotation*, and $\mathbf{H}_{\text{R}} = \mathbf{K}\mathbf{R}\mathbf{K}^{-1}$. Using the focal length ratios $s_x = f'_x/f_x$ and $s_y = f'_y/f_y$, we have:

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix},\ \mathbf{K}' = \begin{bmatrix} f'_x & 0 & c'_x \\ 0 & f'_y & c'_y \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{K}_{\text{rel}} = \begin{bmatrix} s_x & 0 & c'_x - s_x c_x \\ 0 & s_y & c'_y - s_y c_y \\ 0 & 0 & 1 \end{bmatrix}.$$

Therefore, $\mathbf{K}_{\text{rel}}$ is independent of the absolute focal lengths. Conventional image augmentations, including random rotation, scaling, and translation within the image plane, can be interpreted as special cases of the proposed camera augmentation.

We now add lens distortion, for which we define the generic distortion and undistortion function $\mathcal{D}_{\mathbf{K},\boldsymbol{\theta}}(\cdot)$ and $\mathcal{D}^{-1}_{\mathbf{K},\boldsymbol{\theta}}(\cdot)$. These functions depend on the intrinsic parameters in $\mathbf{K}$ and the distortion coefficients $\boldsymbol{\theta}$, which typically model radial and tangential distortions. For the perspective function $\Pi([q_1, q_2, q_3]^\top) = [q_1/q_3, q_2/q_3]^\top$ and the homogenisation function $\Gamma([q_1, q_2]^\top) = [q_1, q_2, 1]^\top$, the complete transformation is:

$$\mathbf{p}' \sim \mathcal{D}_{\mathbf{K}',\boldsymbol{\theta}'}\left(\Pi\left(\mathbf{K}_{\text{rel}}\mathbf{H}_{\text{R}}\Gamma\left(\mathcal{D}^{-1}_{\mathbf{K},\boldsymbol{\theta}}(\mathbf{p})\right)\right)\right). \tag{3.5}$$

In performing camera augmentation, $\mathbf{K}_{\text{rel}}$, $\mathbf{R}$ and $\boldsymbol{\theta}'$ are drawn from probabilistic distributions that model realistic variations in camera parameters in the target domain. These distributions act as a bridge between the known camera configuration in the available dataset and the perturbed configuration in the augmented setup, allowing camera augmentation to be performed without re-rendering. The use with stereo laparoscopes is discussed in the next sections, where we measure that lens distortion is negligible for the public datasets at hand. We thus use the following

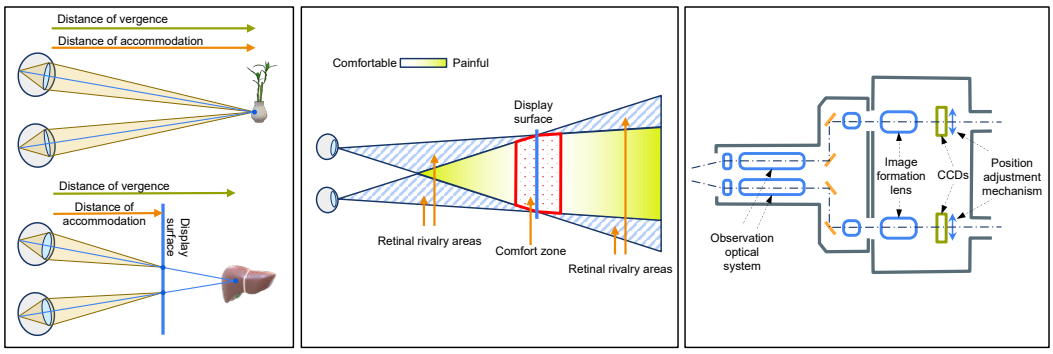


Figure 3.6: Left: VAC; In natural viewing, vergence and accommodation coincide (top). In stereo displays, vergence targets the virtual object while accommodation stays on the display plane (bottom). Middle: Combined effects of binocular rivalry and vergence–accommodation conflict, highlighting the range where both are minimised. Right: CCD position adjustment mechanism in a stereo laparoscope bringing the virtual image into the visual comfort zone.

distortion-free simplification of Eq. 3.5:

$$\mathbf{p}' \sim \mathbf{K}_{\text{rel}}\mathbf{H}_{\text{R}}\mathbf{p}. \tag{3.6}$$

3.3.4 Stereo Perception and the Design of Stereo Laparoscopes

We analyse the perceptual constraints underlying stereoscopic vision and the design principles of stereo laparoscopes. This analysis provides the bases for modelling realistic variations in intrinsic and extrinsic camera parameters, as described in section 2. While the design considerations discussed here are representative of modern systems, such as those used in da Vinci platforms and further examined in this work, they may not be strictly followed by all manufacturers.

Overview. Depth perception in humans may be enabled by displaying stereo images, should some perceptual constraints be satisfied. A stereo laparoscope is a dual-channel imaging system that captures synchronised image pairs for such displays. Its base components, shown in figure 3.6, are two parallel optical pathways with relay lenses, fibre-based illumination, and twin sensors. The sensors are factory-calibrated to provide left and right views for depth perception via a binocular eyepiece or 3D display. The surgeon's visual discomfort, including fatigue and eye strain, and quality of depth perception, directly depend on the extent to which the stereo imaging pipeline's design satisfies the perceptual constraints [Banks et al., 2012]. These constraints and design choices can be gathered in two groups, which we describe next, along with an experimental verification.

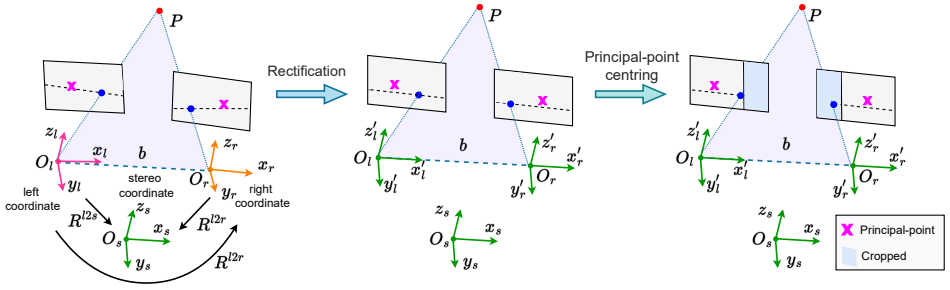


Figure 3.7: Geometry of stereo camera pairs in unrectified (left), rectified (middle), and principal-point aligned (right) configurations. The principal point shift and scan-line misalignment are intentionally exaggerated for visualisation purposes. In practice, the applied perturbations are much smaller and preserve sufficient overlap between the two views.

Geometric alignment. In stereo vision, depth perception requires precise alignment of the left and right images. In an ideal stereo configuration, correspondences exhibit purely horizontal displacements. This condition, referred to as rectification in computer vision, ensures that epipolar lines are horizontal and aligned. The careful design and factory-based placement adjustment of the imaging lenses and sensors allow modern stereoscopic laparoscopes to produce near-rectified image pairs.

To characterise how closely real systems approximate this ideal configuration, we quantify the residual misalignment using the stereo camera parameters provided in the public datasets SCARED [Allan et al., 2021b], StereoMIS [Hayoz et al., 2023], and RIS2017 [Allan et al., 2019], acquired with *da Vinci* systems. We used the extrinsic parameters to estimate three rotations around each camera's optical centre. These rotations are computed relative to the stereo coordinate system shown in figure 3.7, defined with the X-axis aligned with the baseline and taken as minimal with respect to the left camera. Statistics -mean and standard deviation- of the sampled angles are given in table 3.6.

Overall, the rotations are tightly concentrated around zero and remain well below 1°. Their consistency across datasets indicates that different laparoscopes exhibit similarly small deviations from the rectified design. In particular, the small rotations around the Y-axis in table 3.6 show that the display planes follow a planar design without toe-in.

We further use the intrinsic parameters to quantify lens distortion, which is known to impair rectification. The measured residual distortion is consistently below half a pixel; given the image resolution of 1280×1024, this confirms a near distortion-free design. Finally, the principal points of both cameras must be aligned; this aspect is discussed below.

Vergence-accommodation conflict and binocular rivalry. Depth perception in stereo vision involves two physiological processes: vergence, which controls the

Table 3.6: Statistics on camera pose deviations from the ideal rectified configuration for the left (top row) and right (bottom row) cameras, for the three Euler rotation angles, in degrees.

	R_x^{l2s}		R_y^{l2s}		R_z^{l2s}		R_x^{r2s}		R_y^{r2s}		R_z^{r2s}	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
SCARED	0.0000	0.0001	-0.0325	0.0129	0.0904	0.3597	-0.0014	0.0053	0.0013	0.0266	0.0910	0.3602
StereoMIS	-0.0006	0.0006	0.3673	0.0032	-0.1907	0.1812	-0.0020	0.0008	0.3799	0.0187	-0.1887	0.1809
RIS2017	-0.0004	0.0014	-0.3386	0.1652	0.2637	0.4680	0.0002	0.0076	-0.2744	0.1640	0.2639	0.4681
SCARED+StMIS	-0.0001	0.0004	0.0564	0.1666	0.0280	0.3487	-0.0016	0.0047	0.0854	0.1594	0.0289	0.3489

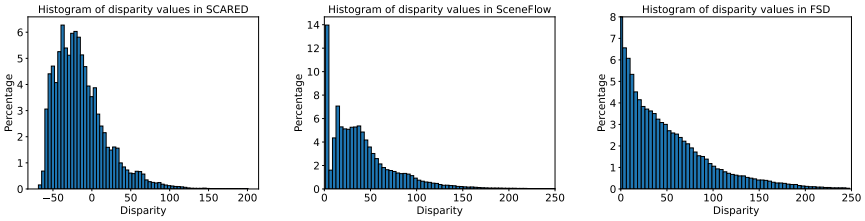


Figure 3.8: Disparity distributions (in px) for the SCARED, SceneFlow, and FSD datasets.

convergence of the eyes to maintain single vision, and accommodation, which adjusts the focus of the eye. In natural viewing conditions, these processes occur at the same distance. However, in stereoscopic displays, a mismatch arises: vergence is driven by the perceived depth of virtual objects, while accommodation remains fixed on the display plane. This mismatch is known as the Vergence–Accommodation Conflict (VAC) [Kramida, 2015]. In addition, binocular rivalry may occur when the images presented to the left and right eyes are inconsistent, for example due to differences in field of view. This effect becomes more pronounced when virtual objects are perceived close to the viewer. Figure 3.6 illustrates the range of depths where the combined effects of VAC and binocular rivalry are minimised.

To mitigate these perceptual issues, stereo laparoscopes are designed such that the perceived depth of the scene falls within a comfort zone. This is typically achieved by adjusting the relative position of the sensors during factory calibration, as described in patented systems [Akui et al., 1996, Zhao et al., 2019, Breidenthal et al., 2000]. In practice, this adjustment shifts the principal points of the cameras horizontally, effectively placing the perceived depth around a typical surgical working distance (approximately 5 cm).

From a geometric perspective, this design choice introduces a systematic offset of the principal points away from the image centre. This creates a domain gap between stereo laparoscope images and conventional stereo datasets, where principal points are typically centred. One direct consequence is that disparities in laparoscopic images can take both positive and negative values. In contrast, most state-of-the-art stereo matching methods are trained on datasets with only positive disparities,

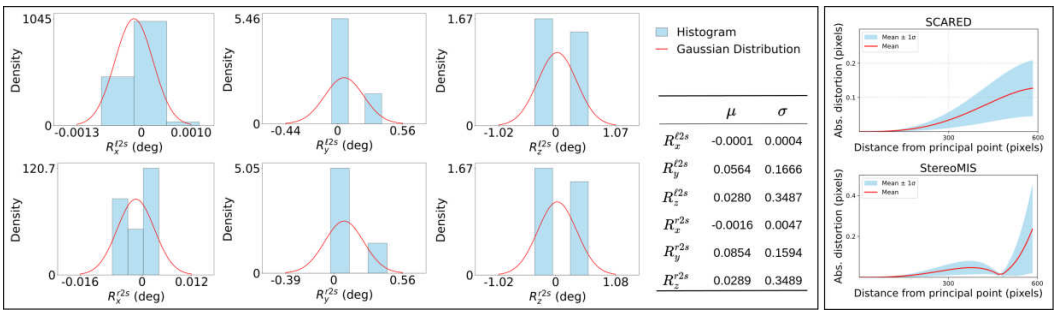


Figure 3.9: Left: Distribution of camera pose deviations from the ideal rectified configuration for the left (top row) and right (bottom row) cameras across the SCARED and StereoMIS datasets, for the three Euler rotation angles. Right: Radial distortion profiles for the StereoMIS (top) and SCARED (bottom) datasets, as functions of pixel distance from the principal point.

such as SceneFlow [Mayer et al., 2016] and FoundationStereo Dataset (FSD) [Wen et al., 2025]. Real surgical datasets, such as SCARED [Allan et al., 2021b], exhibit both positive and negative disparities, highlighting this discrepancy, as illustrated in figure 3.8.

3.3.5 Camera Augmentation in Stereo Laparoscopes

We use camera augmentation from section 3.3.3 to address uncalibrated stereo matching in MIS. The proposed method, termed CATS, is used to retrain existing stereo matching architectures on camera-augmented datasets that simulate realistic imaging imperfections. We retrained RAFTStereo and IGEV++ using the original training protocol, hyperparameters, and datasets, with as sole modification the use of CATS. Both models are trained on the SceneFlow dataset [Mayer et al., 2016], which contains over 39K stereo pairs with a resolution of 960×540 pixels, rendered from a wide range of 3D scenes. The SceneFlow dataset uses ideal stereo camera geometry, including perfect rectification, centred principal points, and the absence of lens distortion, differing markedly from the real MIS conditions.

We use the analysis of MIS stereo camera imperfections from section 3.3.4 (deviations in rectification and principal point misalignment), to generate camera augmentations. This augmentation strategy encourages the network to learn a stereo matching process adapted to uncalibrated MIS images. Concretely, for each of the left and right cameras, we sample three Euler angles, ϕ, θ, and ψ, from Gaussian distributions reflecting the measured perturbation statistics, shown in table 3.6:

$$\phi_c \sim \mathcal{N}(\mu_{c,\phi}, \sigma^2_{c,\phi}), \quad \theta_c \sim \mathcal{N}(\mu_{c,\theta}, \sigma^2_{c,\theta}), \quad \psi_c \sim \mathcal{N}(\mu_{c,\psi}, \sigma^2_{c,\psi}), \quad c \in \{\text{left}, \text{right}\}. \tag{3.7}$$

The rotation matrix $\mathbf{R}$ is constructed from the Euler angles (ϕ, θ, ψ) following the Z, Y, X (yaw–pitch–roll) convention as $\mathbf{R}_c = \mathbf{R}_z(\psi_c)\,\mathbf{R}_y(\theta_c)\,\mathbf{R}_x(\phi_c)$, $c \in \{\text{left}, \text{right}\}$.

The corresponding homography induced by the rotational perturbation for each camera is computed as:

$$\mathbf{H}_{\mathrm{R},c} = \mathbf{K}\,\mathbf{R}_c\,\mathbf{K}^{-1}, \quad \forall c \in \{\text{left}, \text{right}\}, \tag{3.8}$$

where $\mathbf{K}$ is the intrinsic calibration matrix from the original dataset. We then sample $\mathbf{K}_{\text{rel}}$ for the left and right cameras. Since random scaling is already included in the standard data augmentation pipelines of the reference methods, additional variation in scaling factors, s_x and s_y, would be redundant, and we use $s_x = s_y = 1$ for both cameras. Furthermore, based on observations from section 3.3.4, we use $c'_y = c_y$ for both cameras, as the vertical principal point displacement was found to be negligible. The dominant parameter affecting $\mathbf{K}_{\text{rel}}$ is the horizontal offset between the principal points of the left and right cameras. Accordingly, we set $c'_{x,\text{left}} = c_{x,\text{left}}$ and sample the right camera offset, $\Delta x_{\text{right}} = c'_{x,\text{right}} - c_{x,\text{right}}$ from a uniform distribution estimated from the empirical distribution given in figure 3.8 as $\Delta x_{\text{right}} \sim \mathcal{U}(-100, 0)$. The uniform distribution is a pragmatic choice ensuring coverage of the observed empirical range and keeping the sampling simple. The use of non-parametric distributions would greatly increase complexity. This distribution strikes a balance between positive and negative disparities during training, improving robustness across varying stereo configurations. We arrive at the following sampled relative intrinsic matrices:

$$\mathbf{K}_{\text{rel,left}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{K}_{\text{rel,right}} = \begin{bmatrix} 1 & 0 & \Delta x_{\text{right}} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.9}$$

Finally, according to observations from section 3.3.4, the effects of lens distortion are negligible within the analysed MIS datasets; hence, distortion is omitted in CATS.

Importantly, we apply the same homography as for the left image to the disparity ground-truth and then adjust all disparity values by adding Δx_{right}.

3.3.6 Experimental Setup

Reference models. We consider three state-of-the-art stereo matching architectures: RAFTStereo [Lipson et al., 2021], IGEV++ [Xu et al., 2025], and FoundationStereo [Wen et al., 2025]. All models are used as released, without additional training or fine-tuning, thereby reflecting their generalisation capabilities in unseen domains. RAFTStereo formulates disparity estimation as an iterative refinement process based on a correlation pyramid and recurrent updates. IGEV++ extends this paradigm by incorporating adaptive geometric encodings within the cost volume, enabling improved handling of varying disparity ranges. FoundationStereo is a recent large-scale foundation model trained on the FSD dataset, comprising approximately one million synthetic stereo pairs with realistic rendering effects. Despite their differences, all three models are trained on synthetic datasets (e.g., SceneFlow and FSD) generated under idealised conditions, including perfect rectification, centred principal points, and distortion-free imaging. As a consequence, they implicitly

assume positive disparities and stable camera geometry, which contrasts with real MIS imaging conditions.

For comparison with models specifically designed for MIS, we included MSDESIS [Psychogyios et al., 2022] and MCF-SMSIS [Wu et al., 2024], representing the state-of-the-art in MIS stereo matching. These two models were both fine-tuned on the SCARED training set [Allan et al., 2021b].

We used RAFT [Teed and Deng, 2020] as a representative optical flow method, taking the horizontal component of the predicted optical flow as an estimate of disparity. Finally, we used DUSt3R [Wang et al., 2024b] as a representative general multi-view reconstruction method.

CATS models. We retrain RAFTStereo and IGEV++ using the proposed CATS framework, resulting in the CATS-RAFTStereo and CATS-IGEV++ models. Training is performed using the same datasets, protocols, and hyperparameters as in the original implementations, with the sole modification being the introduction of camera augmentation.

CATS augments each stereo pair through homographic transformations to simulate realistic deviations from ideal stereo laparoscope configurations. These transformations are parameterised by rotations sampled from Gaussian distributions fitted to measured angular deviations, and by principal point shifts sampled from empirically derived distributions. This procedure enables the generation of geometrically consistent image pairs that mimic uncalibrated acquisition conditions, while preserving the underlying scene structure. FoundationStereo is not retrained due to the unavailability of its training pipeline; however, the proposed augmentation strategy is fully compatible with its data-driven design.

Experimental protocol. Quantitative evaluation is conducted on the SCARED dataset, while qualitative results are presented on SCARED, StereoMIS, and an in-house dataset acquired during hepatectomy procedures, collected under ethical approval (IRB00008526-2019-CE58, CPP Sud-Est VI, Clermont-Ferrand, France).

For quantitative evaluation, we use the keyframes of the SCARED dataset, excluding sequences exhibiting inconsistencies between visual observations and kinematic ground-truth, as reported in [Allan et al., 2021b]. In addition, datasets 4 and 5 are discarded due to known annotation errors.

We evaluate all models under two distinct settings. In the *calibrated* scenario, intrinsic and extrinsic parameters are used to rectify stereo pairs and align principal points prior to inference. In the *uncalibrated* scenario, models are directly applied to the original, unrectified image pairs, without any preprocessing.

In the calibrated scenario, we used the available intrinsic and extrinsic parameters to rectify the stereo pairs and centre the principal points before applying the stereo matching models. In the uncalibrated scenario, we applied stereo matching directly on the original stereo image pairs. Disparity accuracy is measured using the End-Point Error (EPE) and Bad-3.0, while depth estimation is evaluated using Mean

Table 3.7: Uncalibrated scenario evaluation over original images from the SCARED dataset. Depth values are computed from predicted disparity using static calibration parameters for reference only. In the DUSt3R method, the calibration parameters are used to find disparity values from the predicted depth.

	Disparity metrics (px)		Depth metrics (mm)			
	EPE ↓	Bad 3.0 ↓	MAE ↓	RMSE ↓	AbsRel ↓	$\delta_{0.25}$ ↑
RAFTStereo	244.37	80.08	43.95	48.69	0.55	0.23
IGEV++	186.49	79.79	39.75	46.03	0.49	0.24
FoundationStereo	331.41	83.60	48.50	52.05	0.62	0.20
MSDESIS	88.25	99.96	38.06	40.42	0.53	0.001
MCF-SMSIS	119.01	97.78	43.38	46.51	0.59	0.03
DUSt3R (GT calib.)	67.81	100	232.78	233.30	4.17	0.00
DUSt3R (DUSt3R calib.)	687.97	100	232.78	233.30	4.17	0.00
RAFT	3.00	27.79	2.85	4.98	0.039	0.81
CATS-RAFTStereo	1.42	6.36	**1.33**	**2.25**	**0.01**	**0.97**
CATS-IGEV++	**1.41**	**6.32**	**1.33**	2.31	**0.01**	**0.97**

Absolute Error (MAE), RMSE, AbsRel, and $\delta_{0.25}$, as explained in section 3.2.7. In the uncalibrated settings, depth values are derived from predicted disparities using known calibration parameters for reference purposes only. For DUSt3R, disparity is computed from predicted depth using either ground-truth or internally estimated calibration parameters.

3.3.7 Experimental Results

Table 3.8: Calibrated scenario evaluation on undistorted, rectified, and principal point–aligned images from the SCARED dataset.

	Disparity metrics (px)		Depth metrics (mm)			
	EPE ↓	Bad 3.0 ↓	MAE ↓	RMSE ↓	AbsRel ↓	$\delta_{0.25}$ ↑
RAFTStereo	1.21	5.67	1.01	2.32	**0.014**	0.979
IGEV++	1.23	5.56	1.03	2.43	0.015	0.980
FoundationStereo	**1.17**	**5.10**	**0.97**	**2.30**	**0.014**	**0.981**
MSDESIS	1.71	9.71	1.39	2.91	0.0200	0.946
MCF-SMSIS	2.08	11.24	1.49	3.16	0.022	0.939
DUSt3R (GT calib.)	71.82	100	231.10	231.64	4.08	0.00
DUSt3R (DUSt3R calib.)	81.63	100	231.10	231.64	4.08	0.00
RAFT	3.68	26.39	2.86	5.60	0.039	0.83
CATS-RAFTStereo	1.23	5.80	1.07	2.59	0.016	0.978
CATS-IGEV++	1.21	5.52	1.01	2.37	0.015	0.980

Quantitative results for the uncalibrated and calibrated scenarios are reported

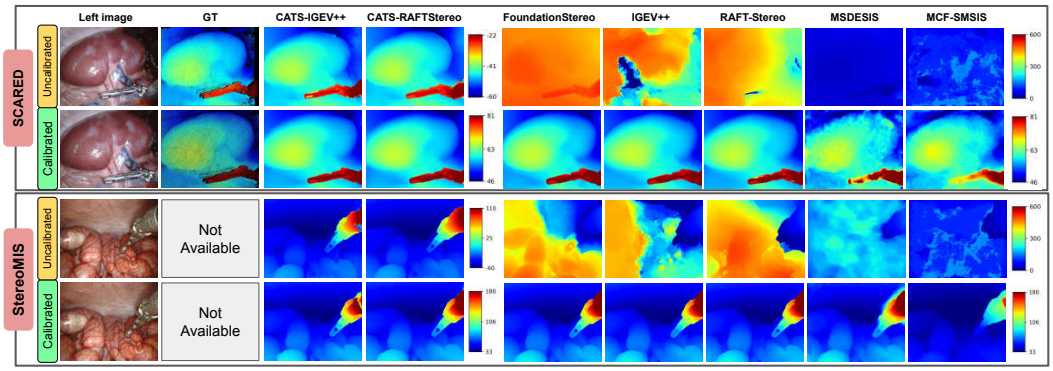


Figure 3.10: Qualitative model comparison in the uncalibrated and calibrated scenarios on samples from the SCARED and StereoMIS datasets.

in tables 3.7 and 3.8, respectively. Several key observations can be made.

First, as shown in table 3.7, all reference and MIS-specific models exhibit systematic failures in the uncalibrated setting, with large disparity and depth errors. RAFTStereo, IGEV++, and FoundationStereo yield EPEs exceeding 180 pixels, and MIS-specific models, MSDESIS and MCF-SMSIS, perform only marginally better with EPEs of 88.25 pixels and 119.01 pixels, respectively, still considered failures. DUSt3R exhibits an EPE of 67.81 pixels when using the camera parameters provided by the dataset, and 687.97 pixels when using its own estimated parameters, both of which are far from acceptable. RAFT achieves a substantially lower EPE of 3.00 pixels, which is nonetheless more than twice as large as the CATS methods. The depth metrics confirm the failures, with the lowest MAE value at 38 mm, indicating poor geometric consistency.

Second, the CATS-trained models perform successfully on the same uncalibrated scenario, achieving EPEs of 1.42 pixels for CATS-RAFTStereo and 1.41 pixels for CATS-IGEV++. Depth metrics further confirm this improvement with an MAE of about 1.33 mm and an AbsRel of about 0.01, demonstrating that CATS successfully enables reliable stereo matching directly from uncalibrated MIS stereo pairs.

Third, the performance of CATS-trained models in the uncalibrated scenario (EPEs of 1.42 and 1.41 pixels for CATS-RAFTStereo and CATS-IGEV++), when compared to the best-performing method in the calibrated setting in Table 3.8 (EPE of 1.17 pixels for FoundationStereo), underscores the effectiveness of CATS. Remarkably, CATS-trained models achieve nearly equivalent performance despite the absence of camera calibration.

Fourth, under the calibrated setting, the same models achieve comparable performance. CATS-RAFTStereo and CATS-IGEV++ reach EPEs of 1.23 and 1.21 pixels, which are on par with their reference counterparts, with RAFTStereo at 1.21 pixels and IGEV++ at 1.23 pixels. This shows that CATS does not compromise performance when calibration is available. Fifth, in the calibrated scenario, CATS-RAFTStereo and CATS-IGEV++ perform nearly as well as FoundationStereo, the

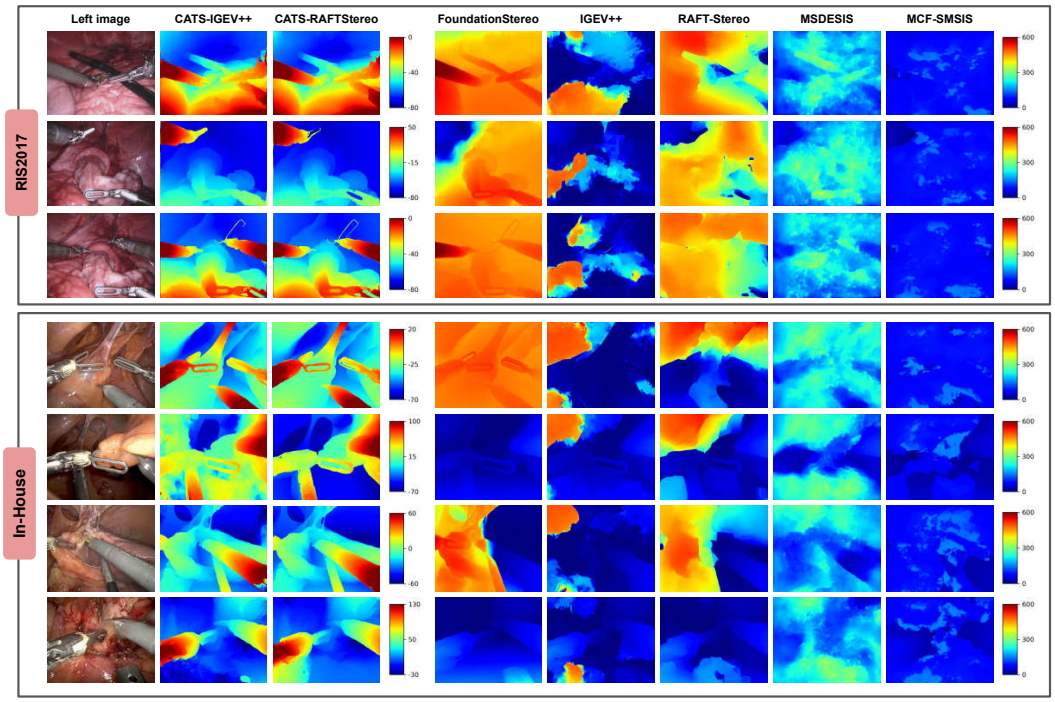


Figure 3.11: Qualitative model comparison in the uncalibrated scenario on samples from the RIS2017 and in-house datasets.

best-performing model, with EPEs marginally higher by 0.04 and 0.06 pixels.

Qualitative results in figure 3.10 and 3.11 illustrate these findings. In the uncalibrated scenario, reference and MIS-specific models produced distorted and inconsistent disparity maps, whereas the proposed CATS models successfully recovered fine structural details and coherent patterns. In the calibrated scenario, all models performed well, including the CATS variants, confirming that CATS remains fully compatible with calibrated conditions and does not degrade performance.

Furthermore, the CATS models demonstrated consistent performance across all tested datasets acquired with different da Vinci system models, highlighting their generalisability. Beyond its strong performance on the three public datasets collected on animal models, CATS produced high-quality, structurally coherent disparity maps on the in-house data of human hepatectomy procedures. This demonstrates the method's robustness and translational potential, supporting its use in real-world clinical applications such as computer-assisted interventions and intraoperative guidance.

3.3.8 Discussion

Camera augmentation effectively bridges the gap between idealised synthetic datasets and the real imaging conditions encountered in MIS, where calibration is often uncertain or unavailable. By introducing controlled geometric perturbations

of camera parameters during training, the proposed approach enables large-scale stereo matching models to adapt to the specific characteristics of surgical imaging.

The CATS framework demonstrates that accurate and robust stereo matching can be achieved directly from uncalibrated image pairs, while maintaining competitive performance in calibrated settings. This capability is grounded in the systematic analysis and modelling of stereo laparoscope geometry, which ensures that the augmentation process reflects the true variability of clinical acquisition conditions.

Beyond the specific case of stereo matching, camera augmentation provides a general framework for incorporating geometric variability into data-driven models. In doing so, it reduces the dependence on precise calibration and promotes the development of methods that are more robust, generalisable, and suitable for clinical deployment.

An important question may arise regarding the interpretation of the recovered geometry under uncalibrated conditions. In conventional stereo vision, disparity is defined under the assumption of rectified image pairs. In practice, stereo laparoscopes produce images that are only approximately rectified, as discussed in section 2. In this near-rectified configuration, vertical disparities remain negligible, allowing horizontal displacements to serve as reliable proxies for disparity. The relationship between disparity and depth can be expressed as $d = \frac{fb}{z} + (c_x^L - c_x^R)$, where f denotes the focal length, b the baseline, and c_x^L, c_x^R the horizontal coordinates of the principal points of the left and right cameras, respectively. This formulation highlights that, in the absence of calibration, depth can only be recovered up to an unknown scale and offset, yielding relative rather than metric reconstruction.

Possible directions for future work include extending the evaluation of CATS to a broader range of stereo surgical systems and incorporating explicit modelling of optical distortion, thereby further improving realism and applicability in diverse intraoperative scenarios.

Chapter 4

Improving Keypoint Matching under Strong Perspective Distortions via Conformal Flattening

---

Contents

4.1 Introduction and Problem Statement

Keypoint matching is a central component of many computer vision pipelines, underpinning tasks such as 3D reconstruction, tracking, and registration. It consists in establishing correspondences between image locations that represent the same physical point across different views. Despite decades of research, achieving reliable matching in real-world conditions remains challenging. Variations in viewpoint, illumination, and scene geometry can significantly alter the appearance of local image regions, making correspondences difficult to establish.

Among these factors, perspective distortion plays a particularly critical role. Specifically, when the camera undergoes large viewpoint changes, the resulting projective transformation induces non-isotropic geometric distortions that cannot be well approximated by local affine models. This perspective distortion warps the images across viewpoints; this warping is both global and local, and arises for both rigid and deformable scenes. As a consequence, the assumption underlying most keypoint detection and description methods —namely that the projective transformation can be sufficiently well approximated locally by an affine transformation— breaks down, leading to reduced repeatability and degraded matching performance.

A natural strategy to address this limitation is to correct perspective distortions prior to matching by warping the image. However, existing warping approaches typically rely on restrictive assumptions, such as scene planarity or developability, which are often violated in general scenes, particularly in surgical settings involving complex, non-planar geometries.

In this chapter, we address the problem of improving keypoint matching under strong perspective distortions on non-planar surfaces. To this end, we propose CPC, an image warping framework based on conformal flattening, a method commonly used in computer graphics for texture mapping. CPC exploits the underlying 3D geometry of the observed surface to construct a mapping that brings the image into an approximately fronto-parallel configuration. The resulting image is virtual: it does not correspond to any physically attainable camera view (a 3D surface cannot, in general, be seen fronto-parallel everywhere simultaneously), but instead provides a locally fronto-parallel representation across the entire surface. This mapping preserves local angular structure while explicitly controlling image resampling. As a result, perspective distortions are reduced, yielding an image representation that is more consistent with the assumptions of standard keypoint matching methods.

In the following, we review existing approaches to keypoint matching under perspective distortions and their limitations, introduce conformal parameterisation as the geometric foundation of our method, and derive the CPC warp. We demonstrate experimentally that CPC improves matching performance and benefits downstream

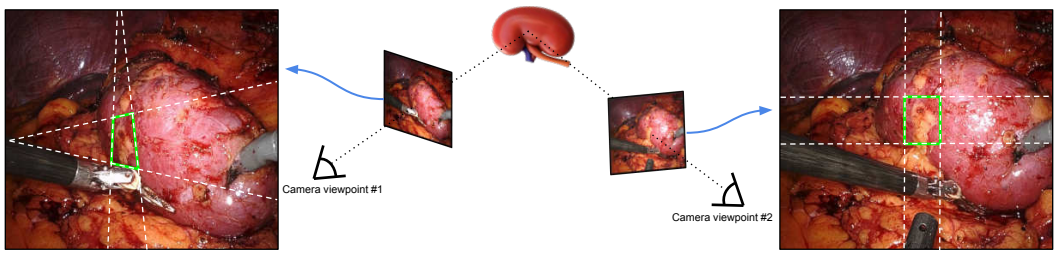


Figure 4.1: Illustration of perspective distortion in MIS. Edges of the selected rectangular patch on the kidney that are parallel in the right image—intersecting at infinity—appear to converge in the left image, forming vanishing points. Source: Author.

tasks such as keyframe matching and 3D reconstruction.

4.2 Keypoint Matching under Perspective Distortions

To motivate the proposed framework, we first revisit the foundations of keypoint matching and examine why perspective distortion constitutes a major failure mode. We begin with the standard detection-description-matching pipeline and the notions of covariant detection and invariant description. We then review existing approaches designed to improve robustness to perspective distortions. This analysis reveals a gap between the geometric complexity of real scenes and the transformation models effectively handled by current methods, motivating the development of a dedicated perspective correction framework.

4.2.1 Keypoint Detection, Description and Matching

Keypoint matching is a fundamental building block in computer vision, enabling the estimation of correspondences across images acquired under varying imaging conditions. It underpins a wide range of applications, including wide-baseline stereo matching [Pritchett and Zisserman, 1998], multi-view reconstruction [Hartley and Zisserman, 2003], image retrieval [Arandjelović and Zisserman, 2012], visual data mining [Simoff et al., 2008], robot localisation [Mur-Artal et al., 2015] and visual servoing [Chaumette and Hutchinson, 2006], panoramic stitching [Brown and Lowe, 2007], and object categorisation [Lazebnik et al., 2006]. In surgical vision, keypoint detection and matching constitute a fundamental component of computer-assisted navigation systems, supporting tasks such as multi-view organ reconstruction [Maier-Hein et al., 2013] and organ-centred camera pose estimation [Grasa et al., 2013].

The standard keypoint matching pipeline consists of three main stages: detection, description, and matching. It is worth noting that recent approaches do not always follow this three-stage pipeline explicitly, but instead directly estimate correspondences in an end-to-end manner [Sun et al., 2021]. Nevertheless, these methods

remain subject to the same fundamental challenges arising from geometric and photometric variations.

Detection aims to identify salient and repeatable local image patches or regions, typically corresponding to regions with high local variation in intensity, such as corners, blobs, or textured patterns. These regions are expected to be stable under changes in imaging conditions.

Description associates each detected region with a compact numerical representation, computed from local image intensities. This descriptor encodes the appearance of the region in a way that facilitates comparison across images.

Matching establishes correspondences by comparing descriptors across images, typically using nearest-neighbour search in a high-dimensional feature space, often followed by outlier rejection through geometric verification using robust estimators such as Random Sample Consensus (RANSAC) [Chandelon et al., 2023], and by additional filtering strategies such as ratio tests or mutual consistency checks.

The effectiveness of this pipeline relies on two key properties: repeatability, i.e. the consistent detection of the same physical points across images, and distinctiveness, i.e. the ability of descriptors to discriminate between different points. These two properties jointly determine the reliability of correspondence estimation: repeatability without distinctiveness leads to ambiguous matches, while distinctiveness without repeatability leads to missing correspondences.

While conceptually simple, this pipeline relies on the implicit assumption that corresponding local image patches remain sufficiently similar across images. In practice, this assumption is challenged by both photometric and geometric variations.

Photometric variations, such as illumination changes, specularities, and sensor noise, can alter the appearance of local patches. These effects can often be mitigated through normalisation strategies, which transform local patches or descriptors into a canonical representation before comparison. This includes operations such as intensity rescaling, contrast normalisation, or gradient-based representations, which reduce sensitivity to global illumination changes while preserving discriminative information.

In contrast, geometric variations are inherently more challenging. They arise from changes in camera viewpoint and scene geometry, inducing local image deformations that depend on surface shape and perspective projection. Most classical keypoint matching methods implicitly assume that these deformations can be locally approximated by similarity or affine transformations, which is valid only under small viewpoint changes or locally planar surfaces. Geometric normalisation is therefore often applied, for instance by aligning patches to a canonical orientation and scale using a dominant gradient direction and characteristic scale.

However, under strong perspective distortions, this assumption breaks down. The induced transformations are no longer well approximated by affine models, leading to inconsistencies in both detection and description. As a result, corresponding regions may not be detected at the same locations, and their descriptors may become incompatible, ultimately degrading matching performance.

This limitation is particularly critical in MIS, where close-range imaging, curved anatomical surfaces, and large viewpoint changes induce significant perspective effects. To better visualise how perspective effects alter the local geometric structure of image patches, two corresponding regions on the kidney surface are highlighted using green polygons in figure 4.1. For a rectangular image patch selected in the right image, the corresponding region in the left image is no longer rectangular due to perspective distortion.

4.2.2 Covariant Detection and Invariant Description

The geometric requirement for reliable keypoint matching is that detected regions correspond to the same physical surface patches across different viewpoints. Since images are projections of a 3D scene, changes in viewpoint induce transformations of these surface patches in the image domain. A keypoint detection and description framework must therefore account for these transformations to ensure consistent correspondence.

Consider two images related by a geometric transformation induced by a change in camera viewpoint. Ideally, a region detected in one image should correspond to the projection of the same 3D surface patch in the other image. This requirement leads to the notion of *covariant detection*, whereby the detection process adapts consistently to geometric transformations. In other words, if an image undergoes a transformation, the detected regions should transform accordingly, preserving their correspondence structure.

In contrast, descriptors are designed to be *invariant*, meaning that they produce similar representations for corresponding regions despite geometric and photometric variations. Invariance is typically achieved through normalisation procedures, such as aligning the region orientation, scaling to a canonical size, or applying affine normalisation before descriptor computation. The goal is to map different appearances of the same physical region to a common representation in descriptor space, while maintaining discriminative power across different regions.

The interplay between covariance and invariance is central to keypoint matching. Detection ensures geometric consistency of the region support, while description ensures appearance consistency. When both properties are satisfied, corresponding regions yield similar descriptors and can be reliably matched.

In practice, most classical detectors and descriptors achieve covariance and invariance only with respect to restricted classes of transformations, which limits their robustness under strong perspective effects. This limitation is reviewed in detail in the next section.

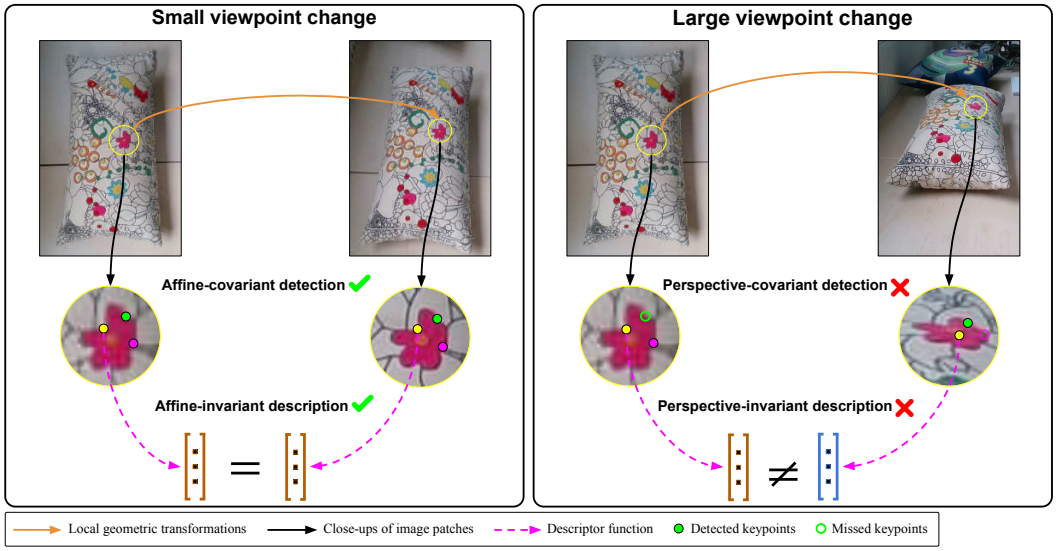


Figure 4.2: Concepts of covariant detection and invariant description. Left: With a small viewpoint change, the local transformation is well approximated by an affine transformation, allowing the keypoints detected in both images to precisely match in position and description. Right: With a large viewpoint change, the local transformation includes a non-negligible perspective component, causing the keypoints detected in both images to slightly differ in position (yellow point) or to be missed (green and pink points), and their descriptors to differ. Source: Author.

4.2.3 Existing Approaches to Perspective-Invariant Keypoint Matching

Existing approaches to handle perspective distortions in keypoint matching can be broadly categorised into two main groups: direct methods and warping methods[1].

Direct methods. Direct methods aim to achieve robustness to geometric distortions at the level of detection and description, without explicitly modifying the image. Existing direct methods involve three levels depending on the class of geometric transformations they can cope with.

Earlier methods such as the Harris method, are limited to invariance under translation and rotation. The entry-level transformation class in modern methods is similarity transformations, including scale and rotation. The most well-known method for the similarity is SIFT [Lowe, 2004], which uses an image pyramid to cope with scale. Although SIFT has shown robustness against moderate perspective distortions, experimental results indicate that strong perspective distortions negatively impact its performance [Mikolajczyk and Schmid, 2004, Mikolajczyk et al.,

[1]Here, *direct methods* refers to approaches that achieve robustness at the level of keypoint detection and description, without explicit image warping. This should not be confused with the classical distinction between direct (photometric) and feature-based image matching methods.

2005, Morel and Yu, 2009, Toft et al., 2020, Rodríguez et al., 2020]. Other widely used similarity-invariant methods include ORB [Rublee et al., 2011] and KAZE [Alcantarilla et al., 2012].

The next level of invariance corresponds to affinity transformations, where we find fewer methods. Notably, as the perspectivity can be locally approximated by an affinity [Rodríguez et al., 2020, Mikolajczyk and Schmid, 2004], the methods designed at the affinity level are also able to handle perspective distortions to some extent. The Hessian-Affine method [Mikolajczyk and Schmid, 2004], which uses local affine shape adaptation, remains a strong baseline among classical methods for image retrieval [Tolias and Jégou, 2014] when combined with RootSIFT [Arandjelović and Zisserman, 2012]. Affinity-adapted methods typically start with detecting the keypoints based on the Harris or Hessian detectors. They then estimate an elliptical shape around the detected keypoint regions, based on the second-moment matrix of the local gradients. They finally normalise the elliptic region to a circular one in order to remove local affine distortions. Although affinity-adapted methods show some robustness against perspective distortion, they fail in 20-40% cases and struggle in the presentation of illumination changes, limiting the keypoint repeatability [Mikolajczyk and Schmid, 2004, Mishkin et al., 2015]. A-SIFT [Morel and Yu, 2009] simulates affine image transformations, in order to find a simulated image pair with the largest number of correspondences. It is however computationally expensive and sensitive to other factors including image blur and illumination changes [Tareen and Saleem, 2018, Wu et al., 2013]. To address these shortcomings, learning-based methods were developed to estimate the local affine shapes, which outperform classical affine methods in image matching [Mishkin et al., 2018].

Finally, the last level transformation class is perspectivity, where one only finds recent learning-based methods. Approaches including LIFT [Yi et al., 2016] and SuperPoint [DeTone et al., 2018], together with matching methods, notably SuperGlue [Sarlin et al., 2020], demonstrate robustness against perspective distortions. Additionally, the latest advance, OmniGlue [Jiang et al., 2024], uses foundation models to further enhance matching generalisability. These learning-based matching methods achieve robustness against perspective distortion mostly by training on numerous images with perspective warps and distortions. Although these networks address the perspective distortions to a higher extent, there is ongoing research aimed at further enhancing their performance. In particular, [Wang et al., 2023] leverages the monocular depth predictors to analytically calculate the local curvature of the 3D patches around keypoints and use it to enhance keypoint matching. However, the method needs the depth estimation network and keypoint matcher to be jointly fine-tuned with curvature information. Additionally, it can only be used with learning-based keypoint methods and does not work with classical approaches. It also includes a hyper-parameter that controls the influence of curvature similarity on keypoint matching.

Warping methods. Warping methods attempt to produce a perspective distortion free image, prior to matching. In other words, instead of designing invariant descriptors, these approaches aim to cancel the perspective distortions by producing a virtual image, upon which standard keypoint detection and matching can be directly applied.

To achieve this, some works use external methods to first recover scene geometry, for instance using vanishing points [Criminisi et al., 2000] or repeated structures [Pritts et al., 2014], and use this information to warp the image. However, applying the warp estimated from planar objects to the entire image produces inaccurate results, leading to strong distortions that adversely affect the subsequent matching process [Toft et al., 2020].

Other approaches exploit 3D information obtained from depth sensors or monocular reconstruction techniques. This 3D information may be available from various sources, including 3D imaging devices such as laser scanners and RGB-D sensors, or through MoSDE methods. A typical approach is to use 3D information to detect planes in 3D scenes and find a perspectivity to transform the image into an orthographic projection. The method in [Toft et al., 2020] uses [Godard et al., 2019] to generate a depth-map from a single monocular image and clusters the surface normals into three dominant directions, representing three planar regions. For each region, a perspectivity is estimated to eliminate the corresponding perspective distortion by warping. Keypoints are then detected in the rectified image for subsequent image matching. Although the method is effective in the nominal conditions, it is limited to scenes with a small number of dominant planar objects.

Concerning non-piecewise-planar scenes, the method in [Zeisl et al., 2012] corrects the perspective effect for objects with developable surfaces, i.e., surfaces that can be flattened onto a plane without stretching or tearing, such as cylinders and cones. The method detects these objects in the depth map and unrolls their surfaces into a 2D domain. They first detect developable objects in the depth map and unroll their surfaces into a 2D domain. Keypoint detection is then performed on the resulting *unrolled* images. The method is effective in the nominal conditions, however only applicable to objects with developable surfaces.

Conformal flattening methods have also been explored for perspective correction [Kim et al., 2012, Chandelon et al., 2023]. While they preserve angular structure, existing approaches often rely on fixed-boundary conditions or neglect the impact of image resampling, leading to distortions and degradation of local image features.

Limitations. Despite substantial progress, existing approaches remain insufficient for handling strong perspective distortions in general scenes. Direct methods rely on local invariance assumptions—typically limited to similarity or, at best, affine transformations—and therefore struggle to cope with the non-linear, depth-dependent distortions induced by perspective projection.

Warping methods, on the other hand, attempt to correct these distortions ge-

ometrically, but are often restricted to specific scene structures, such as piecewise planar or developable surfaces. Moreover, many of these approaches do not adequately account for the distortions introduced during image resampling, which can degrade the local image structures essential for reliable keypoint detection and description.

These limitations are particularly critical in MIS, where anatomical surfaces are highly curved, non-developable, and observed under significant viewpoint changes. This setting highlights the need for a geometry-aware correction framework capable of handling general surface geometries while preserving the local image properties required for reliable matching. In particular, conformal surface parameterisation provides a principled approach to transform such surfaces into a planar domain while preserving their local geometric structure. This motivates the introduction of surface parameterisation and conformal flattening in the following section.

4.3 Surface Parameterisation and Conformal Flattening

This section first introduces the general framework of surface parameterisation, highlighting its principles and associated distortions. It then focuses on conformal mappings, which provide a principled approach for preserving local geometric structure and form the basis of the proposed perspective correction framework.

4.3.1 Surface Parametrisation

Surface parameterisation consists in mapping a 3D surface onto a 2D domain. Given a surface embedded in $\mathbb{R}^3$, the objective is to define a bijective mapping to $\mathbb{R}^2$, enabling the representation of surface geometry in a planar domain. This mapping, commonly referred to as *flattening*, is a fundamental tool in geometry processing.

An intuitive example is the representation of the Earth's surface on a 2D map (figure 4.3). The Greek astronomer Claudius Ptolemy (100–168 A.D.) was among the first ones to systematically describe this problem [Berggren and Jones, 2000]. In his work Geography, he explains how to project a sphere onto a flat piece of paper using a system of gridlines [Ptolemy, 1991]. However, since the Earth is approximately spherical, its planar representation inevitably introduces distortions. This problem is formally studied in cartography through map projections, which define transformations designed to preserve selected geometric properties, such as angles or areas, depending on the application.

Beyond cartography, surface parameterisation plays a central role in multiple domains. In computer graphics, it is used for *texture mapping* (UV mapping), where a 3D surface is unwrapped into a 2D domain to allow texture painting (figure 4.3). In geometry processing, it facilitates operations such as remeshing and shape editing. In medical imaging, parameterisation enables the analysis of complex anatomical structures by mapping them to canonical domains, for instance when flattening cortical surfaces.

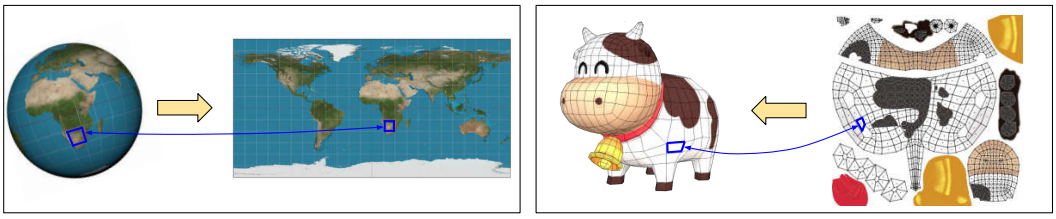


Figure 4.3: Left: Representation of the Earth's 3D surface on a 2D map in cartography. Right: Example of texture mapping (UV mapping) in Blender. The 3D surface is flattened into a collection of 2D patches, referred to as an atlas. This parameterisation defines a bijective mapping (blue arrow) between the surface and the 2D domain, enabling the transfer of texture information onto the 3D geometry. Source: Author.

However, the flattening of general 3D surfaces onto a plane inevitably introduces distortion. This limitation can be intuitively understood by considering the flattening of a curved surface, such as an orange peel, onto a plane: such a surface cannot be mapped without distortion, and compromises are therefore unavoidable. More fundamentally, mappings cannot simultaneously preserve all geometric properties—such as lengths, areas, and angles—when transforming a non-developable surface into a planar domain.

As a result, surface parameterisation methods are typically designed to preserve specific properties while allowing controlled distortion in others. Distance-preserving, or *isometric*, mappings aim to maintain lengths but are generally infeasible for curved surfaces. Area-preserving, or equi-areal, mappings preserve surface area but may significantly distort local shapes. In contrast, angle-preserving, or *conformal*, mappings maintain local geometric structure while allowing variations in scale. While the ideal goal is to achieve an isometric transformation, this is an unreachable goal for a general surface.

The choice of parameterisation is therefore inherently task-dependent. In applications such as texture mapping, preserving local shape is often more important than preserving area, which makes conformal mappings particularly attractive. In this work, we focus on conformal parameterisation, as its ability to preserve local geometry provides a natural foundation for mitigating perspective distortions. The underlying rationale for this property is developed and discussed in section 4.3.3.

4.3.2 Conformal Mapping

A conformal mapping is a transformation that preserves angles. More precisely, it preserves the angle between any two intersecting curves on the surface. As a consequence, the local geometry is preserved up to a similarity transformation, i.e., a combination of translation, rotation, and uniform scaling.

From a theoretical perspective, Riemann's mapping theorem[2] guarantees that any surface homeomorphic to a disk admits a conformal mapping to the plane [Riemann, 1851]. This result provides a strong foundation for the use of conformal parameterisation in practice. Although not immediately intuitive, it establishes the existence of angle-preserving parameterisations for a broad class of surfaces.

From a differential perspective, conformality implies that the Jacobian of the mapping is locally isotropic, meaning that it scales all directions equally at each point. This property ensures that small structures retain their shape, although their size may vary.

While achieving conformal parameterisation is possible theoretically by considering the angles infinitesimally, in practice, surfaces are represented as discrete meshes composed of edges and nodes. Consequently, conformal mappings must be approximated numerically by defining and minimising an appropriate discrete energy.

Several methods have been proposed to compute conformal parameterisations on meshes, and comprehensive surveys can be found in [Floater and Hormann, 2005, Hormann et al., 2007]. Among these, the Least Squares Conformal Mapping (LSCM) method [Lévy et al., 2002] is one of the most widely used and is implemented in many 3D geometry processing tools, including Blender, CGAL, and plugins for Autodesk 3ds Max and Maya.

The original formulation of LSCM [Lévy et al., 2002] enforces conformality by solving a linear system that approximates the Cauchy–Riemann equations in the least-squares sense. The formulation relies on complex-valued functions, where conformality is expressed through constraints on their derivatives. While mathematically elegant, this formulation is not always intuitive and can be difficult to interpret geometrically. In addition, LSCM requires fixing the positions of at least two vertices in the parameter domain to eliminate trivial degenerate solutions, thereby introducing artificial constraints that may influence the resulting mapping. To address these limitations, we propose in section 4.4.3 an alternative formulation of conformality based on barycentric coordinates, expressed directly in the real domain.

4.3.3 Relationship Between Perspective Projection and Conformal Mapping

Perspective projection and conformal mapping are fundamentally different transformations. We study their relationship through their local geometric behaviour. Perspective projection is a projective transformation that maps 3D points onto a 2D image plane through the camera model introduced in section 2.2. While preserving line straightness and incidence, it does not preserve angles and lengths. As a result, it introduces anisotropic, depth-dependent distortions: local image patches may undergo non-uniform scaling and shearing depending on their local surface orientation relative to the camera and their distance to it (figure 4.4). This alters

[2]The theorem was established in the doctoral dissertation of Bernhard Riemann (1851), which laid the foundations of conformal mapping theory.

the local similarity structure of image patches and leads to inconsistencies in keypoint detection and descriptor matching, which typically rely on local similarity assumptions.

In contrast, a conformal transformation is defined by its ability to preserve angles locally. The connection between these two transformations arises when analysing perspective projection at a local scale. Although perspective projection is globally projective, it can be locally approximated by an affine transformation. These local affinities can be constrained by enforcing conformality, thereby preserving angles and limiting perspective distortions.

In summary, perspective projection introduces distortions that violate local similarity, whereas conformal mappings restore local geometric consistency. This makes them particularly well suited for applications such as keypoint detection and matching, where preserving local structure is essential.

Roughly speaking, if we had a conformal transformation of the 3D object to the 2D domain instead of the projective one arising from the camera model, the output would be more adapted to the task of image matching.

In the following section, we introduce the idea that conformal mappings can be adapted to compensate for perspective distortions in practice.

4.4 Perspective Correction via Conformal Flattening

Up to this point, we have established that perspective projection introduces strong viewpoint-dependent geometric distortions, and that exploiting conformal transformations provides a principled way to mitigate them. However, to effectively benefit from such transformations in image matching, important theoretical and practical limitations must be considered. In this section, we first investigate these limitations, and then introduce an image warp for perspective correction based on conformal transformations, which we term CPC. We further derive two optimisation formulations of this principle, including a convex formulation that leads to a Linear Least Squares (LLS) problem.

4.4.1 Theoretical and Practical Considerations

In order to benefit from conformal transformations in image matching, there are important theoretical and practical limitations that must be considered.

1) Restricted class of conformal mappings in higher dimensions. A fundamental limitation arises from the theory of conformal mappings itself. In 2D, conformal maps are abundant and are typically characterised by complex analytic functions [Blair, 2000]. This richness is formalised by the Riemann Mapping Theorem, which states that any simply connected domain, excluding the entire plane, can be conformally mapped to the unit disk.

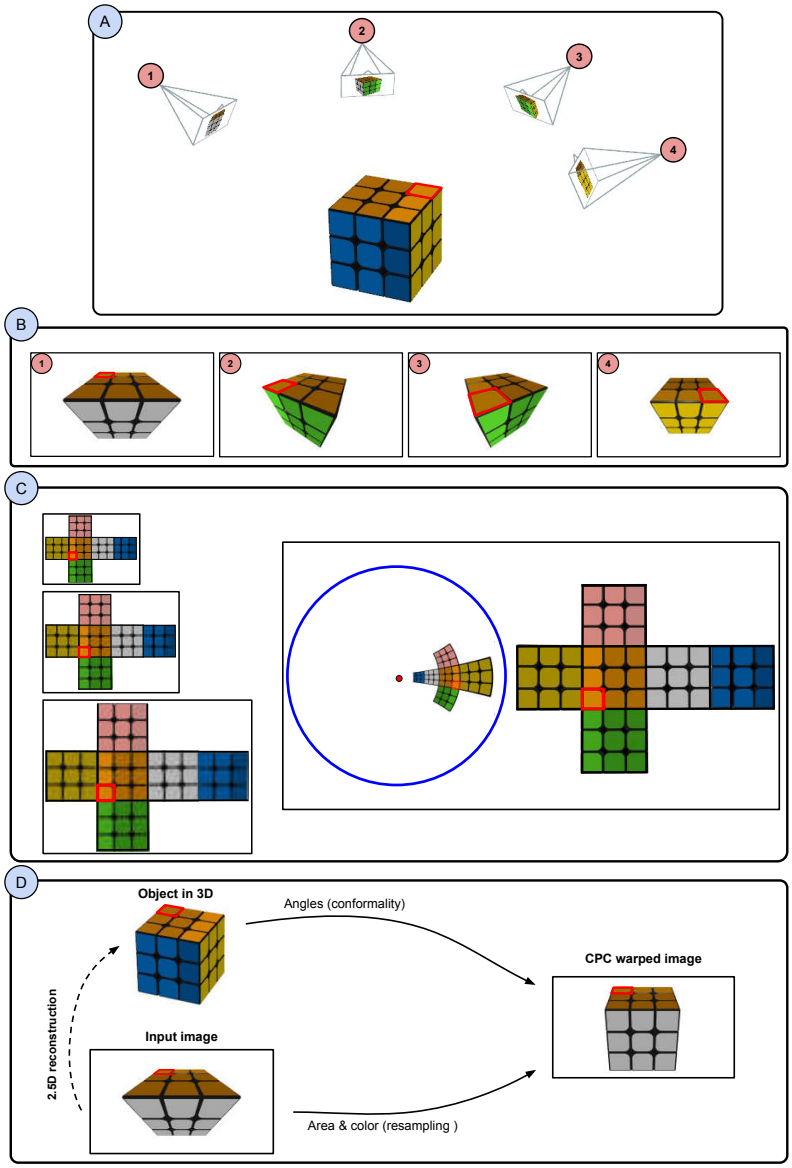


Figure 4.4: Comparison between perspective projection and conformal mapping. A: The scene geometry (a simple cube) and camera configurations. B: Examples of perspective projection of the cube onto the image plane, illustrating the corresponding geometric distortions. The shape of the red polygon varies significantly with the camera viewpoint. C: Conformal flattening of the same cube. The parameterisation preserves local shape (angles), but is not unique: infinitely many mappings exist up to a global similarity transformation, three of which are illustrated (left), along with an example obtained via circular inversion, i.e. a Möbius transformation that maps the plane through inversion with respect to a circle [Blair, 2000]. This transformation preserves angles while altering distances and global structure, further illustrating the non-uniqueness of conformal parameterisations beyond global similarity transformations. D: Proposed CPC image warp. Given an input image and its associated 2.5D reconstruction, the method produces a warped image in which perspective effects are reduced while preserving the image signal as much as possible.

In contrast, the situation is different in 3D and higher dimensions. According to Liouville's Theorem on conformal mappings [Flanders, 1966], any smooth conformal transformation in $\mathbb{R}^n$, for $n \geq 3$, must be a composition of a finite set of elementary transformations: translations, rotations, uniform scalings, and inversions, which together form the class of Möbius transformations.

This theorem severely restricts the space of admissible conformal transformations in 3D. Unlike in 2D, where one can freely design conformal mappings, in 3D the conformal structure is highly constrained. As a consequence it is not possible to arbitrarily correct perspective distortions directly in $\mathbb{R}^3$ using conformal mappings. Instead, one must rely on surface parameterisation, i.e., mapping a surface embedded in $\mathbb{R}^3$ onto a planar domain.

2) Scale ambiguity and sampling distortion. A second major limitation arises from the lack of scale control in conformal transformations, as illustrated in figure 4.4C. While angles are preserved, local scaling may vary significantly. As a result, any image warp derived from such a transformation inherently involves pixel creation (over-sampling) and destruction (under-sampling) [Gluckman and Nayar, 2001], thereby altering the original image signal significantly. This introduces resampling artefacts that degrade the consistency of local image structures, which are essential for reliable keypoint detection and matching. Therefore, it is necessary to explicitly control or regularise the scale behaviour of conformal transformations in practical applications.

4.4.2 General Formulation of Conformal Perspective Correction

We formulate image perspective correction as a minimisation problem that searches for an image warp, defined through a two-term cost function $C : \mathbb{R}^n \to \mathbb{R}$. The first term, $C_{persp} : \mathbb{R}^n \to \mathbb{R}$, quantifies the amount of perspective distortion in the warped image, while the second term, $C_{res} : \mathbb{R}^n \to \mathbb{R}$, quantifies the amount of resampling induced by the warp on the original image. These two terms correspond to objectives that are both necessary to achieve for a successful result, but are generally incompatible. We therefore introduce a hyperparameter $\lambda \in [0, 1]$ to control their respective influence, leading to the total cost:

$$C(p) = \lambda C_{persp}(p) + (1 - \lambda) C_{res}(p). \tag{4.1}$$

In the following we will continue with deriving a LLS formulation for the perspective term C_p, using the theory of conformal mapping.

4.4.3 Deriving a Linear Least-Squares Conformal Formulation

Consider a triangle defined by three non-collinear points, denoted as A, B, C, forming a simplex in $\mathbb{R}^3$. A triangle is thus represented as an ordered triplet of points. The triangle ABC in 3D and its conformally flattened counterpart abc in the plane

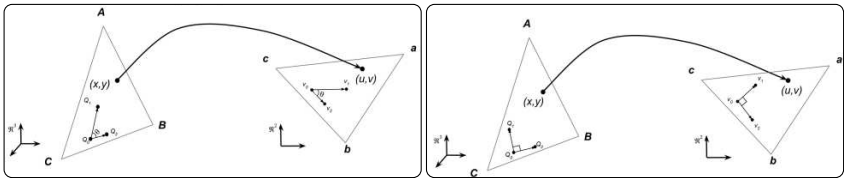


Figure 4.5: Conformal mapping between two triangles, illustrating the local preservation of angles between corresponding bases. Left: General case with arbitrary bases. Right: Special case with orthonormal bases.

are illustrated in figure 4.5. The mapping between these two triangles is affine and conformal if and only if it preserves angles. In practice, this implies that local geometric configurations are related by a similarity transformation, i.e., a translation and a rotation combined with a uniform scaling, while preserving orientation.

To express this constraint, we select three points Q_0, Q_1, and Q_2 within the triangle ABC, defining a local affine basis of the underlying 2D subspace embedded in $\mathbb{R}^3$. For simplicity, we choose these points such that the vectors $\overrightarrow{Q_1Q_0}$ and $\overrightarrow{Q_2Q_0}$ are orthogonal, leading to the constraint:

$$Rot_{90}(\overrightarrow{Q_1Q_0}) = \overrightarrow{Q_2Q_0}. \tag{4.2}$$

This orthogonal formulation is not restrictive, as it can be generalised to arbitrary angles by replacing the 90° rotation with a rotation of angle θ. However, it leads to a simpler and numerically stable formulation.

Under a conformal mapping, the same relation must hold in the parameterisation domain. Denoting by v_i, $i \in 0, 1, 2$, the 2D coordinates of the mapped points, we enforce:

$$Rot_{90}(\overrightarrow{v_1v_0}) = \overrightarrow{v_2v_0}, \tag{4.3}$$

This condition ensures that the flattened triangle is a positively scaled and rotated version of its 3D counterpart, thereby preserving local angles and orientation.

To relate this constraint to the triangle vertices, we express the points Q_i using barycentric coordinates with respect to ABC. Let $[\alpha_j, \beta_j, \gamma_j]$, $j \in 0, 1, 2$ denote these coordinates. The corresponding 2D points v_i are then expressed using the same barycentric weights applied to the unknown vertices a, b, c:

$$v_i = \begin{bmatrix} \alpha_j & 0 & \beta_j & 0 & \gamma_j & 0 \\ 0 & \alpha_j & 0 & \beta_j & 0 & \gamma_j \end{bmatrix} \begin{bmatrix} x_a & y_a & x_b & y_b & x_c & y_c \end{bmatrix}^\top, \tag{4.4}$$

where $x_{a,b,c}$ and $y_{a,b,c}$ denote the unknown 2D coordinates of the flattened triangle.

Substituting this barycentric formulation into the conformality constraint yields a linear relation in the unknown vertex coordinates. This leads to a homogeneous linear system of the form:

$$M_t p_t = 0, \tag{4.5}$$

where p_t stacks the coordinates of the triangle vertices in the parameter domain, and the matrix M_t depends only on the barycentric coordinate differences, and thus solely on the geometry of the triangle in 3D:

$$M_t = \begin{bmatrix} \alpha_2 - \alpha_0 & \alpha_1 - \alpha_0 & \beta_2 - \beta_0 & \beta_1 - \beta_0 & \gamma_2 - \gamma_0 & \gamma_1 - \gamma_0 \\ \alpha_1 - \alpha_0 & \alpha_2 - \alpha_0 & \beta_1 - \beta_0 & \beta_2 - \beta_0 & \gamma_1 - \gamma_0 & \gamma_2 - \gamma_0 \end{bmatrix}. \tag{4.6}$$

The non-trivial solution of this system defines the coordinates of the conformally flattened triangle *abc*, up to a global similarity transformation.

To extend this formulation to the entire surface, we enforce the conformality constraint over all triangles in the mesh. Denoting by T_{3D} the set of triangles, we define the global conformal energy as:

$$C_{conf}(p) = \Sigma_{t \in T_{3D}} \|M_t p_t\|^2. \tag{4.7}$$

This least-squares formulation aggregates local conformal constraints and yields a global optimisation problem over the flattened vertex positions.

The proposed formulation is convex and presents two main advantages over classical LSCM [Lévy et al., 2002]. First, it offers increased generality. While LSCM enforces conformality through orthogonality constraints derived from complex number analysis, the proposed formulation operates directly in the real domain and can preserve angles between arbitrary vector pairs. As such, LSCM appears as a particular case of the proposed framework.

Second, classical LSCM requires fixing the positions of at least two vertices in the parameter domain to eliminate trivial solutions. In our application context, this would introduce undesirable constraints and potentially distort the mapping. In contrast, the proposed formulation does not require such constraints, as additional terms in the overall optimisation (e.g., image-based constraints) naturally regularise the solution and prevent degenerate configurations.

4.4.4 Maximally-Equiareal Formulation of Conformal Perspective Correction

Image resampling is the creation or destruction of pixels occurring during warping. It is therefore intuitively sound to limit resampling by preserving local pixel area. To explicitly control this effect, we enforce local area preservation within the CPC warp. Concretely, we construct a 2D mesh on the image by triangulating its pixel locations, resulting in a grid of 2D triangles T_{2D}. The equiareal formulation penalises deviations between the original triangle areas and their warped counterparts. The corresponding cost is defined as:

$$C_{\text{area}}(p) = \sum_{t \in T_{2D}} (\text{area}(t) - \text{area}(t, p_t))^2 \tag{4.8}$$

This formulation encourages the warp to preserve local image area, thereby limiting both over-sampling and under-sampling effects[3].

The overall CPC objective then becomes:

$$C_{\text{CPC-Equiareal}}(p) = \lambda\, C_{\text{conf}}(p) + (1 - \lambda)\, C_{\text{area}}(p) \tag{4.9}$$

While intuitive and directly related to the notion of resampling, this formulation is non-convex, which makes optimisation challenging and sensitive to initialisation.

4.4.5 Convex Least-Displacement Formulation of Conformal Perspective Correction

To overcome the limitations of the equiareal formulation, we introduce an alternative least-displacement cost, which indirectly controls resampling by penalising pixel displacement. We formulate a cost function which characterises the extent of image resampling by measuring the displacement of pixels placed on a 2D grid within the image, as:

$$C_{disp}(p) = \|P_{init} - p\|^2, \tag{4.10}$$

where P_{init} represents the initial position of the grid. We initialize P_{init} by pruning a regular grid so that it approximately covers the region of interest. This LLS term promotes minimal pixel displacement and therefore limits resampling.

When combined with the conformal term defined in equation (4.7), the full objective becomes:

$$C_{\text{CPC-Disp}}(p) = \lambda \sum_{t \in T_{3D}} \|M_t p_t\|^2 + (1 - \lambda)\|P_{\text{init}} - p\|^2 \tag{4.11}$$

This objective can be rewritten as a single LLS problem:

$$C_{\text{CPC-Disp}}(p) = \left\| \begin{bmatrix} \sqrt{\lambda}\, M \\ \sqrt{1-\lambda}\, I \end{bmatrix} p - \begin{bmatrix} 0 \\ \sqrt{1-\lambda}\, P_{\text{init}} \end{bmatrix} \right\|^2 \tag{4.12}$$

which is convex and admits a unique global minimiser. The solution can be obtained efficiently using the Singular Value Decomposition (SVD) of the system matrix, yielding a stable and reliable estimate of the flattened vertex positions.

[3] It is worth noting that, in general, the combination of conformality and equiareal constraints implies an isometric transformation. However, this does not hold in the present setting. In CPC the conformality constraint is defined with respect to the underlying 3D surface, while the equiareal constraint is imposed in the 2D image domain. Since these two constraints are expressed in different spaces from two different shapes, their combination does not imply isometry of the resulting warp. Consequently, the transformation cannot, for instance, be reduced to a rigid or As-Rigid-As-Possible (ARAP) method [Sorkine et al., 2007], and more general deformations must be considered.

4.4.6 Method Pipeline

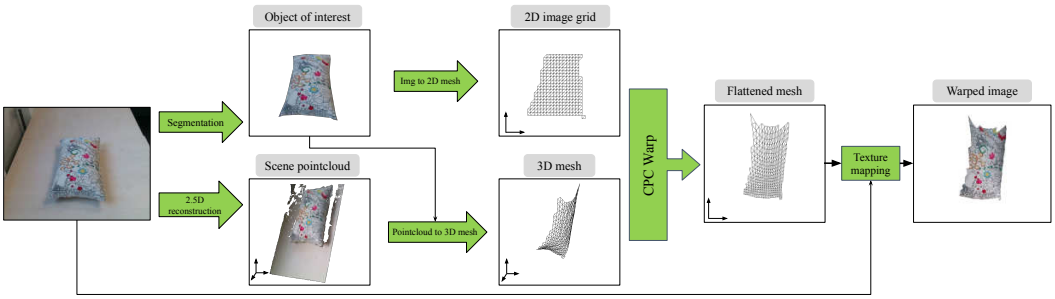


Figure 4.6: The CPC pipeline. Given an input image, the process begins by segmenting the region of interest, here the pillow, and reconstructing its 2.5D geometry. Based on this segmentation, two meshes are generated: a 2D image grid and a 3D mesh of the object. These serve as inputs to the CPC warp (equation 4.9 or equation 4.12), which flattens the 3D mesh while jointly minimising perspective distortion and image resampling. Finally, the input image is used as a texture to generate the warped image.

As outlined earlier, CPC is designed to correct perspective distortions in images while limiting resampling artefacts, in order to improve downstream tasks such as keypoint matching. It therefore takes as input an image together with its associated 2.5D reconstruction (either measured by sensors or reconstructed), and produces as output a warped image in which perspective effects are reduced. As illustrated in figure 4.6, it proceeds in the following five steps:

Step 1: 2.5D reconstruction. It begins by creating a 2.5D reconstruction of the scene generated from the input image. When depth information is available from an RGB-D sensor, it is directly used; otherwise, it can be estimated using one of the reconstruction methods discussed in chapter 3.

Step 2: Region of interest extraction. The Region Of Interest (ROI) is segmented in the input image and the corresponding mask is applied to both the image and its associated reconstruction. The segmentation can be performed interactively; however, automatic approaches are increasingly available. In many applications, including surgery, dedicated learning-based models (e.g., U-Net variants [Ronneberger et al., 2015]) enable reliable segmentation of the objects of interest.

Step 3: 2D and 3D mesh triangulation. Two regular triangulated meshes sharing connectivity are constructed. The first is a 2D mesh defined over the image grid corresponding to the object. The second is the 3D mesh of the same object in the scene, preserving the same connectivity.

Step 4: Warp estimation. The image warp is parameterised by the mesh vertices, whose positions are optimised through cost minimisation. The optimisation is performed in two stages. First, an initial solution is obtained by minimising the convex least-displacement cost 4.11, providing a stable and efficient estimate. This

is followed by a refinement step using the equiareal formulation 4.9, which more strongly enforces local area preservation.

Step 5: Image warping. The warped image is obtained by treating the input image as a texture map and mapping it onto the optimised flattened mesh using piecewise affine interpolation. It is important to emphasise that the resulting image is virtual and does not correspond to any physically realisable camera view, as a 3D surface cannot be globally fronto-parallel, but instead admits a locally fronto-parallel representation across its extent.

The warped image produced by CPC can then be used in downstream tasks. In particular, keypoint detection and matching can be performed in the flattened domain, where perspective distortions are largely reduced, leading to improved robustness. The resulting correspondences can subsequently be transferred back to the original image domain.

4.5 Experimental Results

We begin by qualitatively revisiting the motivating example shown in figure 4.2. As discussed, large viewpoint changes violate the assumptions of covariant detection and invariant description underlying reliable keypoint matching. As illustrated in figure 4.7, CPC warp transforms local image patches into an approximately fronto-parallel configuration, making them more suitable for keypoint matching.

Building on this visual intuition, we now conduct a series of experiments to quantitatively evaluate CPC and assess its impact on downstream applications.

4.5.1 Evaluation with a Liver Phantom

	#Crsp. ↑		#C. Crsp. ↑		Precision ↑		Recall ↑	
	easy	*hard*	*easy*	*hard*	*easy*	*hard*	*easy*	*hard*
SIFT+UBCMatcher	49	12	45	2	0.92	0.17	**0.80**	0.06
CPC+SIFT+UBCMatcher	**50**	**24**	**47**	**12**	**0.94**	**0.74**	0.70	**0.32**
ORB+NN	**145**	0	94	0	0.65	0.00	**0.45**	0.00
CPC+ORB+NN	142	**5**	**96**	**2**	**0.68**	**0.40**	0.36	**0.08**
SuperPoint+SuperGlue	**137**	53	**135**	41	**0.99**	0.52	**0.90**	0.41
CPC+SuperPoint+SuperGlue	136	**76**	133	**68**	0.98	**0.65**	0.88	**0.68**
LoFTR	318	206	295	161	0.93	**0.81**	0.64	0.38
CPC+LoFTR	**324**	**343**	**301**	**264**	0.93	0.77	**0.65**	**0.70**

Table 4.1: Comparison between methods with and without using CPC. The results are obtained using the liver phantom shown in figure 4.8. The column "#Crsp." gives the number of correspondences obtained by the matching algorithm. The column "#C. Crsp." gives the number of correspondences validated using the phantom ground-truth pose. For all metrics, higher is better.

We quantitatively evaluate CPC using a 3D-printed and painted liver phantom.

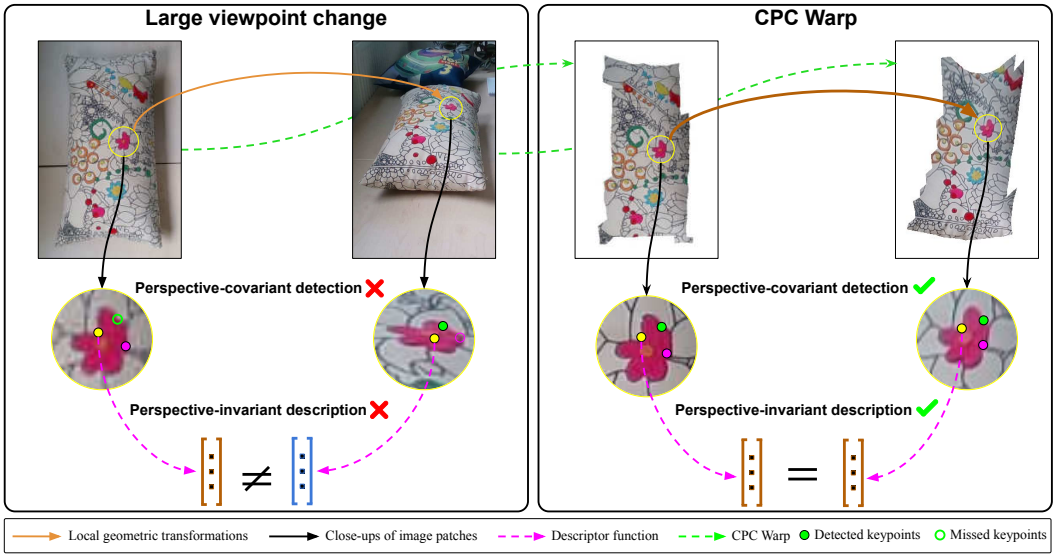


Figure 4.7: Left: The illustrative example from figure 4.2 highlighting the effect of large viewpoint changes, leading to non-covariant keypoint detection and non-invariant description. Right: CPC-warped counterparts, where local patches become fronto-parallel, enabling perspective-covariant detection and perspective-invariant description.

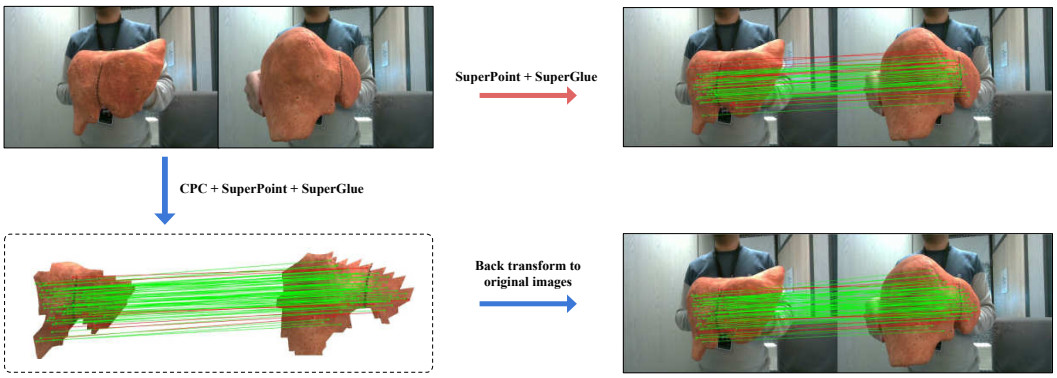


Figure 4.8: Quantitative experiment on a 3D-printed liver phantom. The red arrow illustrates the standard keypoint matching pipeline, using the state-of-the-art SuperPoint detector and SuperGlue for matching. The blue arrows show the proposed CPC framework, which warps the images, performs matching in the flattened domain, and transforms the correspondences back to the original images. Eight markers (small black circles on the phantom surface) are used to facilitate stable pose estimation and provide ground-truth correspondences. Green lines indicate correct correspondences, while red lines denote incorrect matches. In this experiment, CPC increases the number of correct correspondences from 41 to 68, corresponding to a +66% improvement.

The phantom was constructed by first reconstructing a 3D liver model from a patient CT acquired at our hospital under an IRB-approved protocol. Subsequently, randomly spaced carved markers with known 3D locations were added to the surface of the 3D liver model before printing. These markers facilitate ground-truth camera pose estimation. This is particularly important in our experiments, as it enables pixel transfer between images and allows the automatic and reliable assessment of correspondence accuracy. A correspondence is considered valid if its transfer error is below a threshold of 3 pixels. The transfer error is computed as the distance between the keypoint in the second image and the keypoint transferred from the first image to the second, averaged with the distance obtained by reversing the two images to ensure bidirectionality. Finally, after printing, the phantom was painted to reproduce the typical repetitive liver texture, thereby enhancing realism and making image matching more challenging. Images of this phantom were acquired using an Intel RealSense D405 RGB-D camera.

Recall that the motivation behind CPC is that most keypoint detection and matching methods cannot cope well with strong perspective effects. CPC is therefore designed as an optional preprocessing step that can be combined with existing keypoint methods. Consequently, the evaluation should consider representative keypoint detection and matching methods, both with and without CPC.

We use the SuperPoint keypoint detector [DeTone et al., 2018] combined with the SuperGlue matcher [Sarlin et al., 2020], as well as LoFTR [Sun et al., 2021]. These methods represent the state-of-the-art in learning-based matching. We used the main GitHub repository for SuperPoint[4] and SuperGlue[5], and the Kornia[6] implementation [Riba et al., 2020] for LoFTR. In addition, we include SIFT and ORB as representative classical methods, implemented using the VLFeat library[7]. Together, these methods provide a representative set of both learning-based and classical approaches.

Using this setup, we evaluated CPC under two conditions: *easy* and *hard*. As a preliminary step, we performed a sanity check of CPC under easy conditions, characterised by rich texture and minimal viewpoint change. For this evaluation, we selected two consecutive frames of the liver phantom with negligible perspective variation, referred to as the easy frames. In contrast, the hard frames involved significant perspective distortion.

The results, shown in table 4.1 and figure 4.8, demonstrate that CPC provides a significant performance boost in hard conditions across all compared methods, as validated by the ground-truth poses. For the easy frames, no significant degradation is observed in the number of correct correspondences when CPC is used.

[4] https://github.com/rpautrat/SuperPoint
[5] https://github.com/magicleap/superglouepretrainednetwork
[6] https://github.com/kornia/kornia
[7] https://www.vlfeat.org/

4.5.2 Keyframe Matching in Partial Nephrectomy

We conducted an organ tracking experiment on a robot-assisted partial nephrectomy sequence[8]. In this sequence, the surgeon mobilises the kidney with a lateral push to expose it sideways, which is a common surgical gesture during the initial inspection phase. This motion introduces substantial perspective changes from the beginning of the sequence. We selected the first frame of the sequence as the keyframe and evaluated the performance of CPC by matching all subsequent frames to this reference.

Depth estimation was performed using EndoDAC [Cui et al., 2024], and the corresponding point clouds were reconstructed using the camera intrinsic parameters. The results, illustrated in figure 4.9, demonstrate a substantial increase in the number of correspondences when CPC is used.

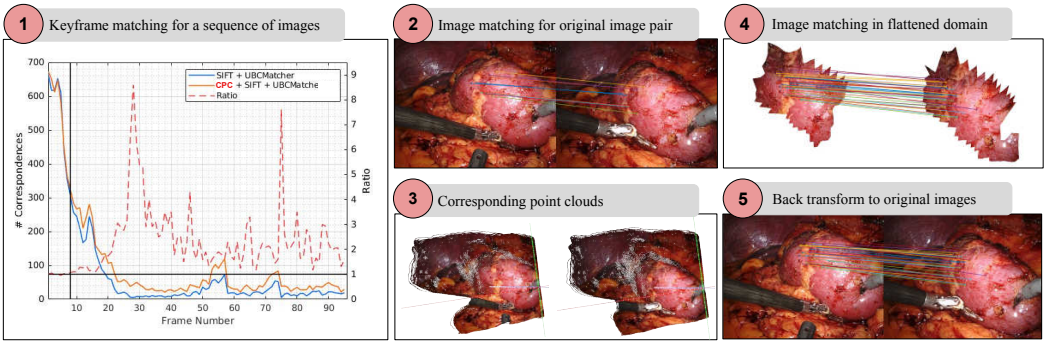


Figure 4.9: Keyframe matching in robot-assisted partial nephrectomy. 1) Keyframe matching results for individual frames of a sequence. The original frame rate of 60 fps was reduced to 5 fps to ease visualisation. The left channel of the stereo camera was used. Frame #1 is chosen as the keyframe and matched to the subsequent frames. The graph shows the number of correspondences of the baseline method without and with CPC, and their ratio. Up to frame #8 (indicated by a vertical black line), the frame difference is minimal and using CPC does not make a difference against the baseline (the horizontal black line represents a ratio of one). Beyond frame #8, as perspective distortions intensify, CPC demonstrates its distortion correction capability and significantly outperforms the baseline. 2) Shows the keyframe on the left and frame #35 on the right, overlaid with the baseline matching result. 3) Shows the monocular point clouds inferred from EndoDAC's depth map and the laparoscope's intrinsic parameters. 4) Shows the matching result in the flattened domain. 5) Shows the result of CPC, as the matched keypoints back-transformed to the original images. Visual inspection did not reveal mismatches, showing that the matches are correct to an excellent extent.

[8]Regarding the data used in this experiment, the patient provided written informed consent in accordance with the 3D Digital Urology / UroCCR project (French network of research on kidney cancer, NCT03293563).

4.5.3 3D Reconstruction

We conducted a 3D reconstruction experiment by integrating CPC with COLMAP [Schonberger and Frahm, 2016, Schönberger et al., 2016], with the objective of evaluating its compatibility with standard SfM pipelines. For comparison, we provided COLMAP with both the keypoints and descriptors obtained using CPC and those obtained without it.

The experiment was performed on the same partial nephrectomy sequence used in section 4.5.2. The reconstructions are illustrated in figure 4.10, and the quantitative results are reported in table 4.2. CPC significantly increases the number of registered images, from 37 to 88, and produces a denser reconstruction, with the number of points increasing from 3,303 to 4,976. In addition, the mean track length improves from 8.43 to 11.23 frames, while the computation time decreases from 3.28 minutes to 1.55 minutes. The mean reprojection error slightly increases from 0.87 to 1.02 pixels, corresponding to a negligible difference of 0.15 pixels. This small increase is marginal compared to the substantial gains in reconstruction completeness and robustness. Overall, these results demonstrate that CPC significantly enhances the performance of SfM, improving both the density and coverage of the reconstructed 3D models.

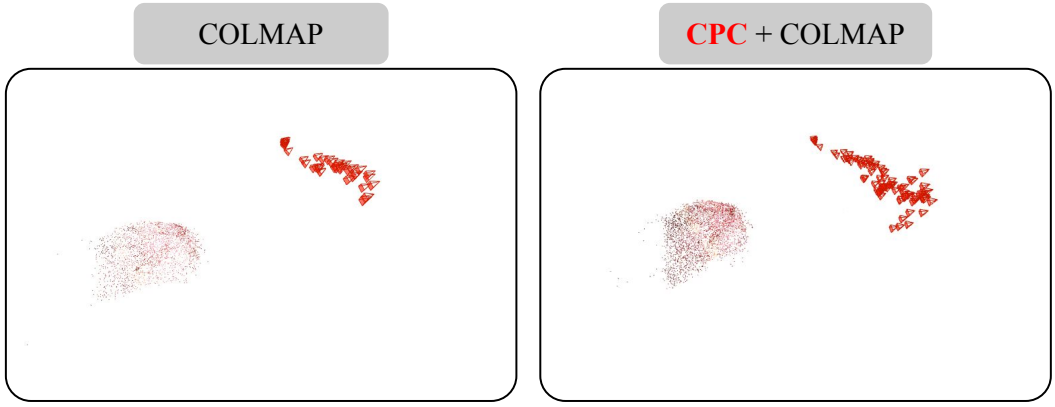


Figure 4.10: Visual comparison of COLMAP reconstruction output. Left: COLMAP used as baseline, without using CPC, Right: COLMAP used with CPC's keypoints.

Method	#Cameras	#Registered Imgs.	#Points	Reproj. Error (px)	Time (min)
COLMAP	37	37/95	3303	**0.87**	3.28
CPC + COLMAP	**88**	**88/95**	**4976**	1.02	**1.55**

Table 4.2: SfM reconstructions obtained by COLMAP without and with CPC.

4.5.4 Automatic Hyperparameter Selection

The hyperparameter $\lambda \in [0, 1]$ controls the trade-off between perspective correction and image resampling. Its choice is critical, as it balances two inherently incompatible objectives: preserving local angles (conformality) from 3D and preserving local areas (equiareality or least displacement) from the original image in the image 2D domain. Due to perspective projection and camera effects, these properties cannot, in general, be simultaneously satisfied.

To illustrate this trade-off, figure 4.11 shows CPC-warped images for five representative values of λ. As λ approaches 1, the CPC warp becomes increasingly conformal but induces higher scale changes thus higher resampling distortion. The trivial spurious solution occurs when $\lambda = 1$. Conversely, as λ approaches 0, the mapping becomes minimally conformal and resampling is minimised. In the limit case $\lambda = 0$, the image remains unchanged. It motivates work on automatic hyperparameter selection.

Ideally, for the application at hand, λ should be chosen to maximise the quality of keypoint matching. However, this objective cannot be optimised directly in practice, as matching quality depends on unknown ground-truth correspondences and is therefore not accessible during warp estimation. However, the guiding principle is to obtain a warp that is as conformal as possible while keeping image resampling under control. Equivalently, we seek the largest value of λ that does not induce excessive area distortion.

To this end, we propose to monitor the amount of image resampling induced by the warp. We define a metric $\delta_{\text{area}}(\lambda)$, which quantifies the relative area change between the original 2D image grid and its warped counterpart. This metric captures how much local pixel areas are distorted during flattening, and therefore provides a direct measure of resampling. Specifically, we define $\delta_{\text{area}}(\lambda)$ as:

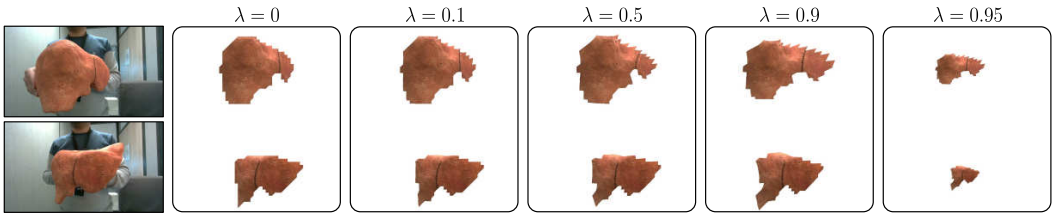


Figure 4.11: Warped images obtained by CPC for five representative values of λ. The original image pair is shown on the left and the warped results on the right. The images illustrate the trade-off between perspective correction and resampling. For $\lambda = 0$, the original image is preserved with no resampling, but perspective distortions remain visible. As λ approaches 1, perspective distortions are progressively corrected and the image appears more fronto-parallel; however, noticeable shrinking occurs, indicating increased resampling. Intermediate values, such as $\lambda = 0.95$, provide a balance between distortion correction and resampling, yielding images that retain their original scale while effectively correcting perspective.

$$\delta_{\text{area}}(\lambda) = \sqrt{\frac{\sum_{t \in T_{2D}} \left(1 - \frac{A^t_{\text{flat}}(\lambda)}{A^t_{\text{init}}}\right)^2}{|T_{2D}|}} \tag{4.13}$$

where A^t_{init} denotes the initial area of triangle t in the 2D image grid, $A^t_{\text{flat}}(\lambda)$ its area after flattening for a given value of λ, and the summation is taken over all triangles $t \in T_{2D}$.

Figure 4.12 b illustrates the behaviour of $\delta_{\text{area}}(\lambda)$ in relation to matching performance. The background heatmap in figure 4.12 b represents the number of correctly matched keypoints, while the curve of $\delta_{\text{area}}(\lambda)$ is overlaid. It can be observed that there is an optimal point on the $\delta_{\text{area}}(\lambda)$ curve where the heatmap achieves its peak value, corresponding to a balance between perspective correction and resampling. Our objective is to automatically identify this optimal λ value. To robustly identify this transition, we model the behaviour of $\delta_{\text{area}}(\lambda)$ using an explicit piecewise function:

$$f(\lambda) = \begin{cases} a_1 \exp\left(-b_1 e^{-c_1 \lambda}\right), & \text{if } \lambda < l, \\ a_2 \exp(b_2(\lambda + c_2)) + d_2, & \text{if } \lambda \geq l \end{cases} \tag{4.14}$$

where the first component follows a Gompertz function and the second an exponential model. To ensure a stable least-squares fitting, we enforce the following constraints:

$$a_1 = \text{median}\left(\delta_{\text{area}}(\lambda)\right) \tag{4.15}$$

$$f(1) = 1 \implies d_2 = 1 - a_2 \exp(b_2(1 + c_2)) \tag{4.16}$$

In addition, continuity conditions (C^0 and C^1) are enforced at the transition point $\lambda = \ell$:

$$a_1 \exp\left(-b_1 e^{-c_1 l}\right) = a_2 \exp(b_2(l + c_2)) + d_2 \tag{4.17}$$

$$a_1 b_1 c_1 e^{-c_1 l} \exp\left(-b_1 e^{-c_1 l}\right) = a_2 b_2 \exp(b_2(l + c_2)) \tag{4.18}$$

Figure 4.12 c compares the measured $\delta_{\text{area}}(\lambda)$ with the fitted model. The selected value $\lambda = \ell$ corresponds to the transition point of this model, providing a trade-off between conformality and resampling.

4.5.5 Effect of Depth Accuracy on CPC's Performance

To analyse the impact of depth estimation quality on CPC, we evaluate several depth estimation methods in the context of intraoperative 3D reconstruction via SfM, as introduced in section 4.5.3. We consider methods that have demonstrated improved accuracy after fine-tuning on surgical datasets, namely AF-SfM Learner [Shao et al., 2022], EndoDAC [Cui et al., 2024], and TRMDSV [Budd and Vercauteren, 2024a].

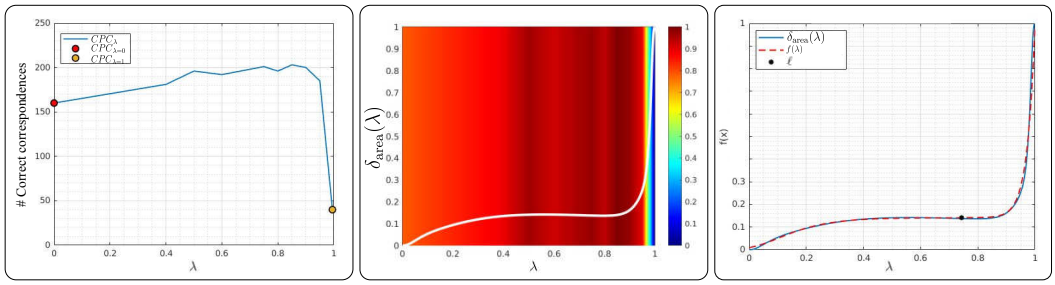


Figure 4.12: Automatic λ selection. a) A pair of images shown in figure 4.8 are flattened using CPC warp for values of λ ranging from 0 to 1. Performance is measured as the number of correct correspondences, validated using camera ground truth. Results are obtained using SuperPoint followed by SuperGlue. Two reference points are highlighted: (1) the red circle at $\lambda = 0$, corresponding to the original images with maximal perspective distortion and minimal resampling; (2) the yellow circle, at a value of λ close to one, corresponding to maximal resampling and the most conformal solution. b) The $\delta_{\text{area}}(\lambda)$ curve (relative area change) is shown in white, overlaid on a heatmap representing the number of correct correspondences. c) The measured $\delta_{\text{area}}(\lambda)$, its fitted model $f(\lambda)$, and the automatic selected trade-off point ℓ are shown together.

Method	#Cameras	#Registered Imgs.	#Points	Reproj. Error (px)	Time (min)
COLMAP	37	37/95	3303	**0.87**	3.28
EndoDAC + CPC + COLMAP	**88**	**88/95**	**4976**	1.02	1.55
AF-SfM + CPC + COLMAP	28	28/95	3108	1.05	3.40
TRMDSV + CPC + COLMAP	84	84/95	4251	0.97	**1.20**

Table 4.3: Impact of the depth estimation network on the final 3D reconstruction. The 3D reconstruction is obtained by SfM via COLMAP without and with CPC.

The evaluation is conducted on the same partial nephrectomy sequence used in sections 4.5.2. The reconstruction statistics are reported in table 4.3. The results indicate that the accuracy of the depth estimation method has a significant impact on the performance of CPC. For example, AF-SfM Learner, which has the lowest reported accuracy among the evaluated methods [Cui et al., 2024, Budd and Vercauteren, 2024a], produces the sparsest reconstruction, with 3,108 points—fewer than the baseline obtained without CPC.

To further analyse this effect, we qualitatively compare the point clouds generated by AF-SfM Learner and EndoDAC, along with the corresponding CPC image warps (figure 4.13). The point clouds are visualised from the same viewpoint to enable direct comparison. Clear differences in the inferred 3D geometry are observed, which lead to different warping behaviours. These discrepancies result in noticeably different perspective corrections. For instance, the close-up example shown in figure 4.13 highlights significant differences between the two methods.

4.6 Discussion

This chapter introduced CPC, an image warp designed to improve keypoint matching under strong perspective distortions on general surfaces. The proposed framework is motivated by the observation that most existing keypoint methods implicitly

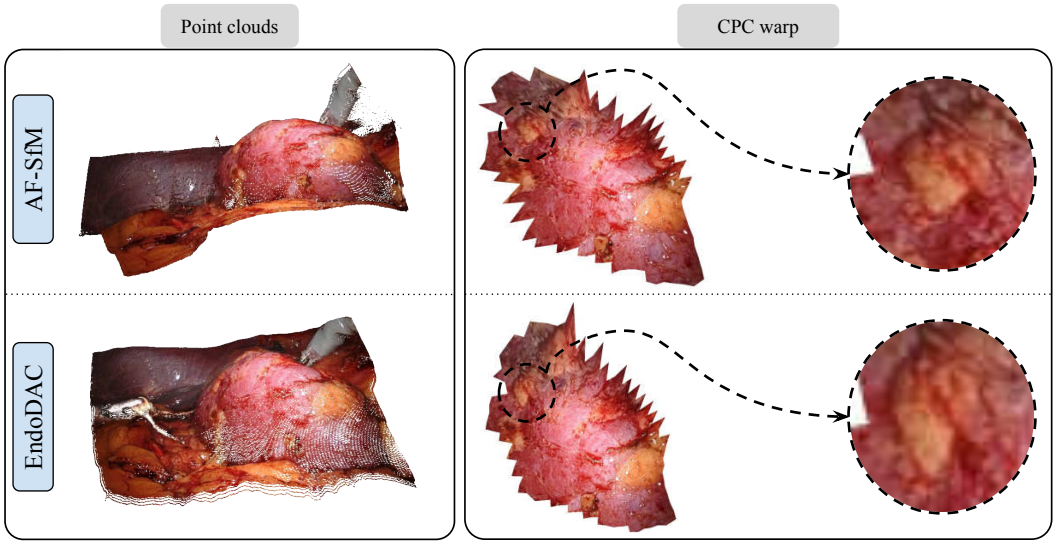


Figure 4.13: Qualitative comparison of the impact of depth estimation quality on CPC. Both 3D point clouds are rendered from the same viewpoint, illustrating significant differences in the inferred geometry. These differences explain the reduced performance of AF-SfM + CPC, as AF-SfM provides less accurate depth estimates [Cui et al., 2024, Budd and Vercauteren, 2024a]. In contrast, EndoDAC + CPC yields more consistent results and improves downstream tasks such as SfM.

rely on local similarity or affine assumptions, which become invalid under large viewpoint changes. By transforming the image into a perspective-reduced representation through conformal flattening while controlling image resampling, CPC aims to restore conditions that are more favourable to reliable matching. The experimental results support this motivation, showing consistent gains in difficult matching scenarios and positive effects on downstream tasks such as keyframe matching and 3D reconstruction.

A first point worth clarifying concerns the role of depth in the proposed framework. At first sight, it may appear paradoxical to use depth estimation in order to improve keypoint matching, given that keypoint matching is itself a component of reconstruction pipelines such as SfM. The rationale is that a reliable keypoint matching would allow one to bootstrap SfM with higher quality keypoint correspondences.

A second point relates to the dependency of the method on depth estimation quality. As highlighted in section 4.13, inaccuracies in 2.5D reconstruction, as well as challenging surgical conditions such as non-uniform illumination, smoke, and bleeding, may affect the reliability of the inferred geometry. These factors can propagate to the warp and degrade its effectiveness, constituting a limitation of the current approach.

A third point concerns the range of downstream applications. While improving matching directly benefits stereo reconstruction, SfM, and visual SLAM, the number of impacted tasks is likely much broader. In particular, it can improve organ tracking, pose estimation and deformable reconstruction methods, to name but a few. Indeed, in contrast to rigid methods such as SfM, where the use of RANSAC brings a strong tolerance to spurious point correspondences, deformable reconstruction methods such as Shape-from-Template (SfT) and Non-Rigid Structure-from-Motion (NRSfM) do not benefit from a global geometric model and cannot use RANSAC. They are thus extremely sensitive to spurious correspondences, which the proposed CPC considerably reduces, as shown in the experiments. Therefore, progress in keypoint matching drives one towards the use of deformable reconstruction which would strongly benefit navigation in soft organs.

A fourth point concerns computational efficiency. To assess runtime, we reimplemented the area-based resampling cost from [Gluckman and Nayar, 2001] within our framework. This cost term is non-convex and requires nonlinear optimisation. On average, this previous term requires 12 seconds of iterative optimisation per image. This is clearly incompatible with real-time processing by several orders of magnitude. In contrast, the proposed CPC framework uses a convex cost function and operates at 120 milliseconds per frame. Regarding the time needed for depth estimation, we evaluated AF-SfM learner Shao et al. [2022], EndoDAC [Cui et al., 2024], and TRMDSV [Budd and Vercauteren, 2024a] models, measuring inference times of 14 milliseconds, 22 milliseconds, and 20 milliseconds, respectively, on an RTX 2080 GPU. Including the depth estimation inference time with CPC, the complete overhead computation time amounts to 145 milliseconds per frame, which aligns

with the real-time processing standards.

In conclusion, CPC corrects perspective distortions by preserving 3D angles and minimising image resampling. The method boosts the performance of existing keypoint matching in the presence of extreme perspective distortions. Importantly, CPC is adaptable to non-planar scenes, has a convex LLS formulation, and can be easily integrated into existing computer-aided surgery systems.

Chapter 5

Topology-adaptive Deformable Registration by Optimisation over Differentiable Simulation

Contents

5.1 Introduction and Problem Statement

Deformable registration is a fundamental problem in computer vision, aiming to establish dense spatial correspondences between images or geometric objects. Unlike rigid registration, deformable registration must account for complex deformations arising for instance from articulated motion or material elasticity. Accurate modelling of these deformations is critical in applications ranging from AR [Collins et al., 2020, Espinel et al., 2021, Pilet et al., 2008], dynamic 3D reconstruction [Newcombe et al., 2015], and shape analysis [Cootes et al., 1995, Rueckert et al., 2003, Modersitzki, 2003].

In the context of MIS, deformable registration plays a central role in AR-guided interventions, enabling the alignment of preoperative 3D models with intraoperative observations to visualise subsurface anatomical structures. However, the deformations encountered in surgical environments pose significant challenges, as soft tissues undergo large, complex, and non-linear deformations driven by physiological processes, external manipulation, and interactions with surgical instruments. In addition, surgical actions such as incision and resection introduce discontinuities that alter the organ structure, corresponding to topology changes. These factors considerably complicate the registration problem and challenge the assumptions underlying most existing approaches.

A key limitation of conventional deformable registration methods lies in their reliance on representations with fixed connectivity, such as meshes. Whether based on optimisation or learning, these approaches implicitly assume topology preservation and are therefore unable to model discontinuities induced by surgical actions. In practice, this often leads to unrealistic deformations, particularly near incision boundaries, where the model is forced to stretch or compress to satisfy constraints instead of allowing separation of material. This limitation reflects a more fundamental issue: the deformation space is typically defined independently of the object's physical behaviour, and does not naturally accommodate topology changes. Addressing this issue requires a formulation in which both geometry and its evolution are governed by physically meaningful principles, while remaining flexible enough to represent discontinuities.

In this chapter, we address the problem of topology-adaptive deformable registration. To this end, we introduce a general framework termed RODS, which formulates registration as an inverse physics problem. Instead of directly estimating a deformation field, the method optimises the initial conditions of a differentiable simulation such that its forward evolution produces a configuration consistent with the observed data. This formulation inherently constrains the solution space to physically plausible deformations. To accommodate topology changes, we adopt a particle-based, mesh-free representation that does not rely on fixed connectivity. We instantiate the proposed framework using the MPM, resulting in a method referred to as Registration by Optimisation over Differentiable Simulation using the Material Point Method (RODS-MPM). By representing the object as a collection of interacting material particles, the approach allows separation and discontinuities to emerge naturally, without requiring explicit remeshing or connectivity updates.

This chapter is organised as follows. Section 5.2 presents the clinical and technical background. Section 5.3 introduces the proposed RODS framework and its instantiation based on the MPM. Section 5.4 reports the experimental results, and section 5.5 discusses the strengths and limitations of the proposed approach.

5.2 Background

In this section, we review the relevant clinical and technical literature. We first introduce the clinical context of surgical resection, highlighting how surgical procedures naturally lead to topology changes. We then discuss deformable registration methods, with a particular focus on their limitations in handling such changes. Next, we review particle-based representations for modelling deformable materials, and finally introduce differentiable simulation as a framework for physics-based optimisation.

5.2.1 Anatomical and Non-Anatomical Resection Approaches

In liver surgery, resection procedures—referred to as *hepatectomy*—are performed to remove pathological tissue, such as tumours, while preserving as much healthy tissue as possible. Depending on the anatomical context, tumour characteristics, and patient-specific constraints, surgeons typically adopt one of two main strategies: anatomical resection or non-anatomical resection [Strasberg et al., 2000] (figure 5.1).

Anatomical resection consists of removing an entire segment by following the intrinsic vascular and biliary organisation of the organ, as defined by the *Couinaud classification* [Couinaud, 1957]. This approach ensures that the tumour and its associated vascular territory are excised together, and is often preferred for oncological safety.

In contrast, non-anatomical resection aims to remove the tumour with a safety margin while preserving as much healthy parenchyma as possible. This strategy does not follow anatomical segment boundaries, but instead relies on the local geometry of the lesion. It is commonly used for small or peripheral tumours, or in patients with limited liver function.

From a deformation perspective, both approaches induce complex organ deformations. Anatomical resections typically result in large-scale but structured deformations aligned with anatomical planes, whereas non-anatomical resections produce more localised and irregular changes. Importantly, in both cases, the surgical act of cutting introduces topology changes, leading to separation of tissue that cannot be captured by conventional deformation models.

These characteristics make liver resection a particularly challenging and clinically

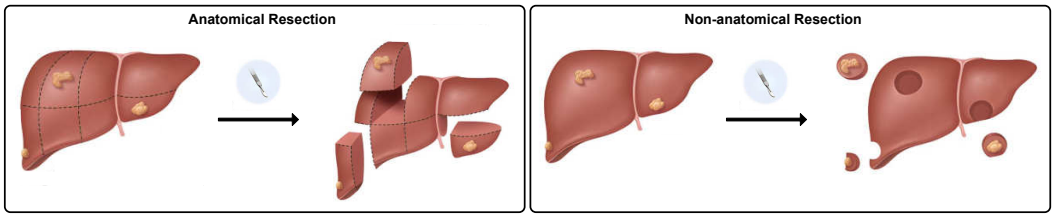


Figure 5.1: Anatomical versus non-anatomical approaches for liver resection. Adapted from [Rengers and Warner, 2023].

relevant scenario for deformable registration with topology changes. In the following, we evaluate the proposed RODS-MPM framework on a clinical case involving a non-anatomical surgical incision.

5.2.2 Deformable Registration

Deformable registration aims to establish dense spatial correspondences between images or geometric objects. In the 3D setting, this is typically formulated as the estimation of a deformation map $\psi : \mathbb{R}^3 \to \mathbb{R}^3$ that aligns a source configuration with a target one.

General formulations. Existing approaches for deformable registration can be broadly categorised into learning-based and optimisation-based methods. Learning-based methods predict deformation fields from data and enable fast inference, but depend strongly on the training distribution and may generalise poorly to unseen scenarios. In contrast, optimisation-based approaches typically formulate registration as a variational problem:

$$\min_{\psi} \; \mathcal{C}_{\mathcal{D}}[\psi] + \lambda \mathcal{C}_{\mathcal{R}}[\psi], \tag{5.1}$$

where $\mathcal{C}_{\mathcal{D}}[\psi]$ is the data fidelity term, which measures how well the transformation fits the data, and $\mathcal{C}_{\mathcal{R}}[\psi]$ is the regularisation term that enforces prior assumptions such as physical plausibility. The hyperparameter λ makes the trade-off between the two.

Regularisation models. A key design choice for deformation models lies in the formulation of the regularisation term. Two main approaches are commonly considered: mathematical models and physics-based models. The former impose purely geometric constraints on the deformation, while the latter constrain it according to the physical behaviour of the object. The most general form, encompassing both mathematical and physics-based approaches can be written as:

$$\mathcal{C}_{\mathcal{R}}[\psi] = \int_{\Omega} \Psi(F(P))\, dP \tag{5.2}$$

where $F(P)$ denotes the local deformation gradient at point P, and Ψ is a scalar-valued density function[1]. Depending on the chosen formulation, Ψ may either represent a purely mathematical regularisation term or a strain energy density

[1] It is worth noting that the majority of the models reported in the literature, whether mathematical or physics-based, are first-order, meaning that they involve only the Jacobian of the deformation map [Sifakis and Barbic, 2012, Modersitzki, 2003]. Higher-order models, such as strain-gradient elasticity [Mindlin, 1965] or curvature regularisers [Fischer and Modersitzki, 2003, 2008, Sotiras et al., 2013] have also been proposed, but first-order formulations remain predominant. We therefore restrict our background review to this class of models.

function derived from continuum mechanics, thereby providing a unified framework that encompasses both classes of approaches.

Mathematical formulations define Ψ as a geometric penalty enforcing desirable properties such as smoothness, isometry or ARAP [Sorkine et al., 2007], and volume preservation [Botsch and Sorkine, 2008]. These formulations are computationally efficient and widely used, but they do not explicitly model the underlying material behaviour, which may lead to unrealistic deformations under complex conditions.

In contrast, physics-based models derive Ψ from continuum mechanics, where deformation is governed by material laws and characterised locally by the deformation gradient [Sifakis and Barbic, 2012].

Constitutive models of materials. In physics-based models, different formulations of $\Psi(F)$, known as *constitutive models*, define the material response by enforcing properties such as volume preservation or rotational invariance, and can be adapted to different types of objects. Linear elasticity provides a simple approximation valid for small deformations, but fails to capture large rotations and nonlinear effects [Sifakis and Barbic, 2012]. To address this limitation, nonlinear models based on the Green strain tensor ensure invariance to rigid transformations and improve physical consistency. More generally, hyperelastic formulations, such as Neo-Hookean [Rivlin, 1948] or Ogden models [Ogden, 1972], express the energy in terms of invariants or principal stretches of F, enabling the modelling of more complex deformations commonly observed in soft tissues. As a result, they are widely used in biomechanical registration [Collins et al., 2020, Labrunie, 2024].

Among these models, we adopt in this work a Neo-Hookean formulation, which provides a simple yet effective description of nonlinear elastic behaviour. This choice is made without loss of generality, as the proposed RODS framework can accommodate a broad class of constitutive models[2]. Neo-Hookean models are nevertheless widely used in the biomechanics literature for soft tissue modelling, as they capture key properties such as near-incompressibility and the ability to undergo large deformations, while remaining computationally tractable [Labrunie, 2024]. The corresponding strain energy density function is given by:

$$\Psi(\mathbf{F}) = \frac{\mu}{2}\left(\mathrm{tr}(\mathbf{F}^T\mathbf{F}) - 3\right) - \mu \ln J + \frac{\lambda}{2}(\ln J)^2, \tag{5.3}$$

where $J = \det(\mathbf{F})$. The Lamé parameters μ and λ are related to the Young's modulus E and Poisson's ratio ν through:

$$\mu = \frac{E}{2(1+\nu)}, \qquad \lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}. \tag{5.4}$$

[2]In particular, the RODS formulation is not tied to a specific constitutive model, and can incorporate more complex material laws or simulation schemes, including those involving known forces or dynamics that are not easily expressed within a classical optimisation framework.

Limitations. Despite the differences between optimisation-based and learning-based methods, as well as between mathematical and physics-based formulations, most deformable registration approaches share a fundamental limitation. They rely on representations with fixed connectivity, such as meshes, which implicitly assume topology preservation. As a result, they cannot naturally model discontinuities such as cutting or separation. In the context of MIS, where surgical actions such as incision or resection occur, this limitation often leads to unrealistic deformations near affected regions, with the material being artificially stretched or compressed rather than allowed to separate in a physically plausible manner.

Only a limited number of works have attempted to address topology changes explicitly, typically by modifying mesh connectivity during the registration process. For instance, [François et al., 2021] proposes an image-based method to handle surgical incisions by adapting mesh connectivity, while [Paulus et al., 2017] introduces local mesh refinements to simulate cutting. Although effective in controlled settings, these approaches introduce additional complexity, require careful maintenance of mesh validity, and remain sensitive to the specific type of topology change. As a result, they are generally not robust and often rely on prior knowledge of the deformation process.

These limitations highlight the need for alternative formulations of deformable registration that do not rely on fixed-connectivity representations and that more naturally integrate physical modelling. We review the relevant background for the former in section 5.2.3, focusing on particle-based methods, and for the latter in section 5.2.4, where we review differentiable simulation.

5.2.3 Particle-Based Methods for Deformation Modelling

In mesh-based methods, such as the Finite Element Method (FEM), the domain $\Omega \subset \mathbb{R}^3$ is discretised into a mesh $\mathcal{T}$ composed of elements (e.g., tetrahedra or hexahedra) with fixed connectivity. However, it relies critically on the assumption that the topology of $\mathcal{T}$ remains unchanged. Under large deformations, this leads to element distortion and potential numerical instability. Furthermore, topology changes (e.g., cutting or separation) require explicit remeshing, which is difficult to integrate into dynamic or optimisation-based frameworks.

These limitations motivate the use of particle-based representations as an alternative to fixed-connectivity discretisations. Instead of representing the domain through explicitly connected elements, the material is sampled by a set of points $\{\mathbf{x}_p\}$, often referred to as *particles* or *material points*[3]. Each particle carries physical quantities such as mass m_p, velocity $\mathbf{v}_p$, and internal variables (e.g., deformation gradient $\mathbf{F}_p$). The deformation is then represented implicitly through the evolution of particle positions.

This representation removes the notion of mesh connectivity and therefore avoids issues related to mesh distortion. Moreover, since particles act as independent sam-

[3]The terms particle and material point will be used interchangeably throughout this chapter.

ples of the material, they can naturally accommodate large deformations and even discontinuities. However, this flexibility introduces new challenges, in particular how to approximate spatial derivatives and solve the governing equations without a predefined mesh structure.

Addressing these challenges requires distinguishing between two fundamental viewpoints for modelling motion: the Lagrangian and Eulerian descriptions.

Lagrangian and Eulerian description of motion. In the Lagrangian description, each particle corresponds to a material point, to which physical quantities are attached and evolve with its motion (figure 5.2, left). The dynamics are described by tracking the position and velocity of individual particles over time. This representation is particularly well suited for capturing deformation history, as it preserves the history of each material point. Methods such as Smoothed Particle Hydrodynamics (SPH) [Gingold and Monaghan, 1977] [4] and Moving Particle Simulation (MPS) [Koshizuka and Oka, 1996] adopt this fully Lagrangian formulation, where interactions between particles are used to approximate differential operators. These approaches are widely used in fluid simulation, particularly for free-surface flows and multiphase phenomena [Monaghan, 2005].

In contrast, the Eulerian description represents physical quantities on a fixed spatial domain (figure 5.2, right). Instead of tracking individual material points, fields such as velocity or density are defined at fixed spatial locations, typically on a grid, while the material flows through this domain. This viewpoint is also widely used in fluid dynamics [Zhu and Bridson, 2005] and naturally avoids mesh distortion, as the discretisation does not move with the material. However, it does not explicitly track individual material points, which complicates the modelling of history-dependent behaviour and the accurate representation of interfaces (e.g., boundaries between different tissues).

These two viewpoints lead to fundamentally different numerical methods. Purely Lagrangian particle methods naturally track material deformation but may suffer from irregular particle distributions and difficulties in computing spatial derivatives. Purely Eulerian grid-based methods offer stable numerical discretisations but struggle to represent large deformations of solid materials and to preserve material coherence. These complementary strengths and limitations motivate the development of hybrid particle–grid methods, which aim to combine the advantages of both representations by coupling Lagrangian particles with Eulerian grid-based computations.

Hybrid Particle–Grid approaches. Hybrid particle–grid methods combine the advantages of Lagrangian and Eulerian descriptions by representing material quantities on particles while performing computations on a background grid. In these

[4] SPH is the oldest mesh-less method, and was initially used for modelling astrophysical phenomena without boundaries such as exploding stars and dust clouds [Monaghan, 2005].

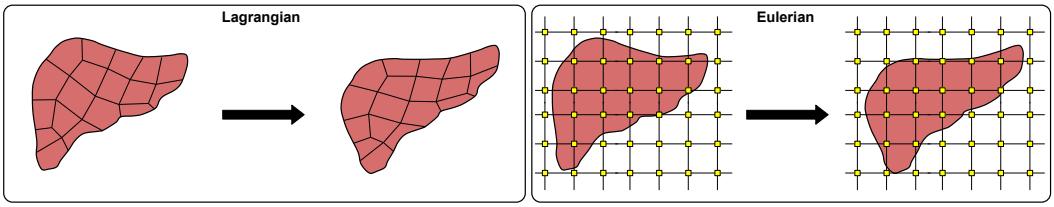


Figure 5.2: Lagrangian versus Eulerian descriptions of deformation. Left: In the Lagrangian description, the grid is attached to the material and deforms along with it, as occurs for example in FEM. Each grid point corresponds to a specific material point, facilitating the modelling of history-dependent behaviour. The object boundary is also well-defined. Right: The Eulerian description uses a fixed grid in space through which the material flows. Mesh distortion never happens. Adapted from [De Vaucorbeil et al., 2020].

approaches, particles carry physical quantities such as mass and velocity, while a grid is used to compute spatial derivatives and solve the governing equations.

Early methods in this family include Particle-In-Cell (PIC) [Harlow, 1962] and Fluid-Implicit Particle (FLIP) [Brackbill et al., 1988]. In these methods, particle quantities are interpolated onto the grid (particle-to-grid transfer), computations are carried out on the grid, and updated quantities are transferred back to the particles (grid-to-particle transfer). The introduction of the grid provides a structured domain for computing gradients, addressing the limitations of purely particle-based methods.

Despite their improvements, both PIC and FLIP were originally developed for fluid simulation and are not directly suited for modelling solid mechanics or complex material behaviour. This limitation motivates the development of more general hybrid formulations capable of handling both fluid- and solid-like materials. The MPM builds upon these ideas by extending hybrid particle–grid formulations to continuum mechanics. It combines the robustness of grid-based computations with a particle-based representation of material state, enabling the simulation of large deformations, complex material behaviour, and topology changes within a unified framework.

Material Point Method. MPM was originally developed as an extension of PIC/FLIP methods for solid mechanics [Sulsky et al., 1994, Nguyen et al., 2023]. MPM has also gained considerable attention in computer graphics, where it has been incorporated into the production pipeline of Walt Disney Animation Studios and applied in feature animations including Frozen, Big Hero 6, and Zootopia [De Vaucorbeil et al., 2020]. The key idea is to combine a Lagrangian representation of material state on particles with an Eulerian grid used as a temporary computational structure.

In MPM, the continuum is discretised into a set of N material points, or parti-

cles, each carrying the full physical state of the material. Each particle p represents a small volume and is associated with physical quantities including mass m^p, volume V^p, position $\mathbf{x}_t^p$, velocity $\mathbf{v}_t^p$, and deformation gradient $\mathbf{F}_t^p$. This particle-based representation enables a Lagrangian description of material motion, while computations are performed on an auxiliary grid. The motion of the continuum is governed by the conservation of mass and momentum. These equations can be written as:

$$\frac{D\rho}{Dt} + \rho\nabla \cdot \mathbf{v} = 0, \qquad \rho\frac{D\mathbf{v}}{Dt} = \nabla \cdot \boldsymbol{\sigma} + \mathbf{f}, \tag{5.5}$$

where ρ denotes the density, $\mathbf{v}$ the velocity field, $\boldsymbol{\sigma}$ the Cauchy stress tensor, and $\mathbf{f}$ external forces.

The key idea of MPM is to alternate between particle and grid representations. At each time step, particles transfer their mass and momentum to a background grid (particle-to-grid transfer), where the governing equations are discretised and solved. This allows the computation of internal forces through the stress term derived from the strain energy density function, consistently with the constitutive model. The momentum at a grid node i is computed as:

$$m_t^i \mathbf{v}_t^i = \sum_p N(\mathbf{x}^i - \mathbf{x}_t^p)\left[m^p \mathbf{v}_t^p + \left(m^p \mathbf{C}_t^p - \frac{4}{(\Delta x)^2}\Delta t V^p \frac{\partial \psi}{\partial \mathbf{F}^p}\mathbf{F}_t^{p\intercal}\right)(\mathbf{x}^i - \mathbf{x}_t^p)\right] + \mathbf{f}_t^i, \tag{5.6}$$

where $N(\cdot)$ denotes the B-spline interpolation kernel, Δx is the grid spacing, Δt the simulation time step, and $\mathbf{f}_t^i$ external forces applied on the grid. After solving for grid velocities, updated quantities are transferred back to the particles (grid-to-particle transfer), enabling the update of particle velocities and positions:

$$\mathbf{v}_{t+1}^k = \sum_i N(\mathbf{x}^i - \mathbf{x}_t^k)\mathbf{v}^i, \quad \mathbf{x}_{t+1}^p = \mathbf{x}_t^p + \Delta t \mathbf{v}_{t+1}^p. \tag{5.7}$$

Interpolation between particles and grid nodes is performed using weighting functions $w_{ip} = N(\mathbf{x}^i - \mathbf{x}_p)$, which determine the interaction between particles and grid nodes. These weights depend on the distance between particles and nodes and are typically defined using B-spline kernels. The choice of interpolation function introduces trade-offs between smoothness, computational efficiency, and numerical stability. In practice, quadratic or cubic B-splines are commonly used to ensure sufficient continuity and avoid instabilities such as cell-crossing artefacts.

A typical MPM time step consists of the following stages (figure 5.3). The particles are assumed to have already been initialised (i.e. each particle has an initial position $\mathbf{x}_p$, initial velocity $\mathbf{v}_p$, mass m_p, and material-related parameters such as μ and λ in the Neo-Hookean model).

1) Particle-to-Grid (P2G) transfer: Particle mass and momentum are transferred to the grid using interpolation weights.

2) Grid-based computation, including the following substeps:

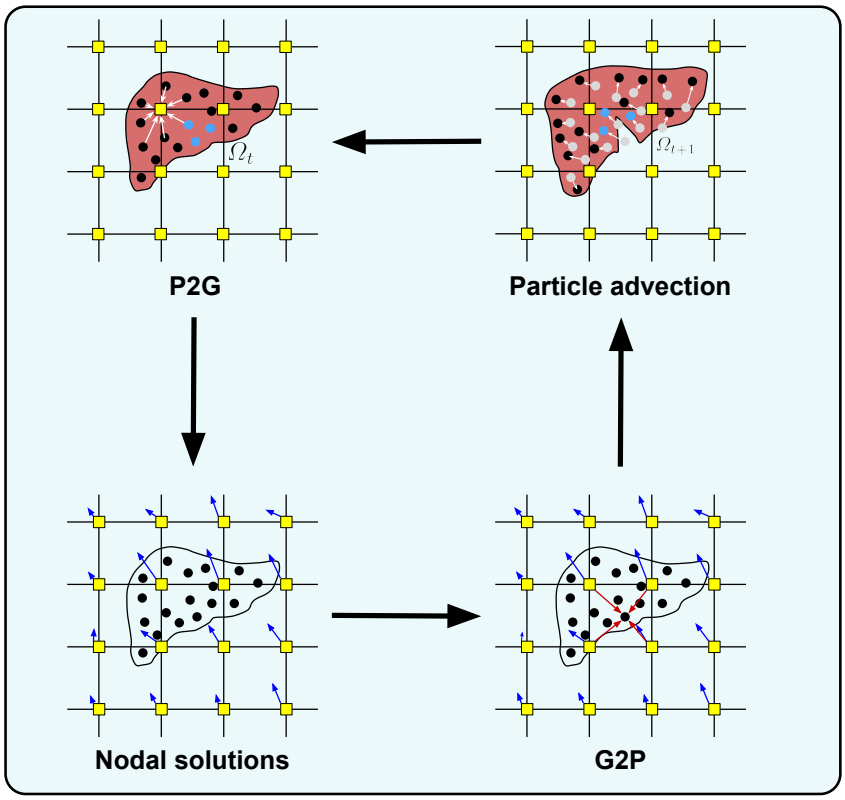


Figure 5.3: Single computational step of MPM.

- *Grid velocity computation:* Grid velocities are computed from the accumulated grid momentum and mass.
- *Active grid node identification:* Grid nodes with non-zero mass are identified as active degrees of freedom, while all other nodes are ignored for computational efficiency.
- *Grid velocity update:* Grid velocities are updated according to the governing equations 5.5.

3) Grid-to-Particle (G2P) transfer: Updated grid velocities are interpolated back to the particles, yielding new particle velocities.
4) Particle advection: Particle positions are updated as $\mathbf{x}_p^{n+1} = \mathbf{x}_p^n + \Delta t \mathbf{v}_p^{n+1}$. The grid is then reset for the next time step.

A key feature of MPM is that the background grid is reconstructed at each time step, avoiding the need for persistent mesh connectivity. This enables robust simulation of large deformations, as particles move independently without causing mesh distortion. Moreover, the absence of fixed connectivity allows the natural handling of topology changes, such as cutting or separation, without requiring explicit remeshing.

However, MPM also presents limitations. The repeated transfer of information between particles and grid may introduce numerical dissipation and interpolation errors. In addition, the accuracy of the method depends on the choice of interpolation kernels and grid resolution, which may increase computational cost. Finally, although MPM avoids mesh distortion, it relies on a background grid, which introduces additional computational overhead compared to purely particle-based approaches.

5.2.4 Differentiable Simulation

Differentiable simulation has recently emerged as a powerful paradigm for modelling physical systems while enabling gradient computation with respect to system inputs, parameters, and initial states [Newbury et al., 2024b, Hu et al., 2019b, Newbury et al., 2024a, Macklin, 2022, Hu et al., 2019a]. Unlike classical simulators, which only predict the forward evolution of a system, differentiable simulators additionally provide derivatives of the simulation outputs with respect to their inputs and parameters. This makes them directly compatible with gradient-based optimisation frameworks providing an efficient framework for solving inverse problems, where unknown quantities must be inferred from observations.

Differentiable simulation enables a wide range of applications, notably system identification [Norouzi-Ghazbi and Janabi-Sharifi, 2020], where physical parameters such as stiffness or damping are estimated; trajectory optimisation [Lutter, 2023, Heiden et al., 2021], where control inputs are optimised to achieve a desired behaviour; morphology optimisation [Hu et al., 2019b], where the physical structure of a system is refined, and policy learning [Newbury et al., 2024b], where control strategies are learned through interaction with the simulator. In all these cases, access to gradients enables efficient optimisation over parameters that would otherwise be difficult to estimate.

Core components of differentiable simulators. A differentiable simulator can be decomposed into four core components: gradient computation, the dynamics model, the contact model, and the numerical integrator [Newbury et al., 2024b]. These components jointly define both the forward simulation and the associated gradient computation, and determine the behaviour, accuracy, and applicability of the simulator.

The first component, gradient computation, is what fundamentally distinguishes differentiable simulators from classical physics engines. It provides derivatives of the simulation outputs with respect to inputs, initial conditions, or control variables. These gradients enable optimisation over the simulation process and allow the simulator to be embedded within learning or inverse problem frameworks. Several approaches exist, including analytical derivation [Song and Boularias, 2020], symbolic differentiation [Newbury et al., 2024b], and automatic differentiation [Baydin et al., 2018]. Among these, automatic differentiation has become the dominant approach

due to its flexibility and scalability.

The second component, the dynamics model, defines the physical laws governing the system. It typically consists of a predefined set of equations derived from kinematic or dynamic principles, for instance continuum mechanics. Depending on the application, the dynamics model may describe rigid-body motion, deformable materials, or fluid behaviour.

The third component is the contact model, which handles interactions between objects, particularly during collision events. Contact modelling is one of the main challenges in differentiable simulation, as it introduces discontinuities in the system dynamics, therefore making gradient computation difficult. Practical approaches often rely on smooth approximations or compliant models to maintain differentiability.

The fourth component is the numerical integrator, which advances the system state over discrete time steps by solving the equations of motion. It uses the computed forces to propagate the dynamics of the system forward in time.

Differentiable simulation engines. Several software frameworks have been developed to support differentiable simulation in practice. DiffTaichi [Hu et al., 2020] provides a high-performance differentiable programming environment based on source-code transformation and supports particle-based, rigid, and soft-body simulations, including MPM. Warp [Macklin, 2022], developed by NVIDIA, enables high-performance simulation and graphics with just-in-time compilation for CPU and GPU execution. Other frameworks include Tiny [Heiden et al., 2021] Differentiable Simulator and Nimble [Werling et al., 2021] for rigid-body dynamics. GradSim [Murthy et al., 2021] also provides end-to-end differentiability from physics to rendering for robotics and perception tasks, and later evolved toward the more general framework design adopted in Warp. While many of these frameworks focus on rigid-body systems, fewer support deformable or topology-changing simulations, which remain more challenging.

Collectively, these simulators highlight key trade-offs between physical fidelity, computational efficiency, and gradient accuracy [Newbury et al., 2024b]. We use DiffTaichi as the computational backbone of our framework due to its support for high-performance particle-based simulation and its native automatic differentiation through time integration. Its data-oriented design allows flexible implementation of constitutive models and simulation pipelines.

The capabilities of differentiable simulation are directly relevant to deformable registration, which can be interpreted as an inverse problem. Instead of directly optimising a deformation field, one can optimise the parameters of a physical simulation such that the simulated configuration matches observed data. In this work, we adopt this perspective and formulate deformable registration as an optimisation problem over a differentiable simulation model.

5.3 Problem Formulation

Registration can be interpreted as an inverse problem in which the goal is to recover unknown variables that generate a deformation aligning a source configuration with target observations. Classical approaches typically optimise the deformation field directly, using regularisation terms to enforce smoothness or physical plausibility. An alternative is to leverage differentiable simulation, where the deformation is generated by a physical forward model and the unknown variables are optimised through gradients of the registration objective.

To this end we propose a novel computational approach for deformable registration termed RODS. Instead of directly estimating deformation fields, RODS formulates deformable registration as an inverse physics problem. Specifically, registration is achieved by optimising the initial state of a differentiable physics simulator such that its forward evolution produces a configuration that aligns with intraoperative observations. This formulation inherently constrains the deformation space to physically plausible states defined by the simulator dynamics, eliminating the need for hand-crafted regularisation terms and ensuring physically consistent deformations. Importantly, RODS is a general framework that can be implemented using any differentiable simulation engine.

In the following, we present the general RODS formulation (section 5.3.1), followed by its MPM-based instantiation (RODS-MPM) for handling topology changes (section 5.3.2).

5.3.1 Registration by Optimisation over Differentiable Simulation

RODS is a computational approach for deformable registration, built around the core component of a differentiable simulator. Let $\mathcal{S}_T$ denote the simulation operator over the time horizon T, which propagates the system state $\mathbf{z}_t$ from time step 0 to time step T, such that $\mathbf{z}_T = \mathcal{S}_T(\mathbf{z}_0)$. The state variables are split into $\mathbf{z}_t = [\mathbf{p}_t, \mathbf{q}_t]^\intercal$. Vector $\mathbf{p}_t$ encodes the object geometry; thus, $\mathbf{p}_0$ and $\mathbf{p}_T$ correspond to the source object and the candidate registered object, respectively. In contrast, $\mathbf{q}_t$ encodes all state variables unknown at $t = 0$, primarily the object dynamics, which are optimised by RODS during the registration process.

The objective of RODS is to find the registered object as the optimal simulated configuration $\mathbf{p}_T^*$ that matches a target configuration $\mathcal{T}$ using a cost function $\mathcal{L}(\mathbf{p}_T, \mathcal{T})$. The principle of RODS is to optimise the initial state $\mathbf{q}_0$, obtaining the final state from simulation as $[\mathbf{p}_T^*, \mathbf{q}_T^*]^\intercal = \mathcal{S}_T(\mathbf{p}_0, \mathbf{q}_0^*)$. This is formulated as the following optimisation problem:

$$\mathbf{q}_0^* = \arg\min_{\mathbf{q}_0} \mathcal{L}(\mathbf{p}_T, \mathcal{T}) \text{ with } [\mathbf{p}_T, \mathbf{q}_T]^\intercal = \mathcal{S}_T(\mathbf{p}_0, \mathbf{q}_0). \tag{5.8}$$

Concretely, in each optimisation iteration, the initial state $[\mathbf{p}_0, \mathbf{q}_0]^\intercal$ is first propagated forward in time to $t = T$ by the simulator. Since the simulator is differentiable

with respect to the state variables, gradients of the cost with respect to the initial state can be computed directly. The initial state is then updated using gradient-based optimisation, using α as the learning rate:

$$\mathbf{q}_0^{n+1} = \mathbf{q}_0^n - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{q}_0}, \tag{5.9}$$

This continues iteratively until the termination condition is satisfied.

As the registered model $\mathbf{p}_T^*$ is the output of a physics simulator, it ensures that the RODS registration output remains physically plausible; in other words the optimisation search space is bound to physically plausible deformations.

5.3.2 RODS-MPM

We instantiate the general RODS framework using a differentiable MPM simulator, such that the forward operator $\mathcal{S}_T(\mathbf{p}_0, \mathbf{q}_0)$ corresponds to the MPM simulation over T time steps.

Object representation and landmarks. The source object is represented using N material particles. Following the RODS formulation, the optimisable system states are selected as initial particle velocities, $\mathbf{q}_0 = [\mathbf{v}_0^i]_{i=1}^N$, while the remaining state variables which encode the object geometry are the initial particle positions $\mathbf{p}_0 = [\mathbf{x}_0^i]_{i=1}^N$.

A set of source landmarks $\mathcal{M}_s = \{\mathbf{x}_0^k\}_{k\in\mathcal{K}}$, with corresponding target landmarks $\mathcal{M}_t = \{\mathbf{y}^k\}_{k\in\mathcal{K}}$, is selected, where $\mathcal{K} \subset \{1, \ldots, N\}$ denotes the set of landmark indices with $|\mathcal{K}| \ll N$. The target landmarks $\mathbf{y}^k$ could be either a point, a line, a curve, or a region.

Registration formulation and optimisation. Following the RODS formulation, registration is achieved by optimising the initial particle state such that the simulated landmark positions match the target landmarks. By optimising the initial particle velocities, the registration problem (5.8) becomes:

$$\mathbf{v}_0^* = \arg\min_{\mathbf{v}_0} \frac{1}{|\mathcal{K}|} \sum_{k\in\mathcal{K}} \text{dist}(\mathbf{x}_T^k, \mathbf{y}^k) \text{ with } [\mathbf{x}_T^k, \mathbf{v}_T^k]^\intercal = \mathcal{S}_T(\mathbf{x}_0, \mathbf{v}_0), \tag{5.10}$$

and where $\text{dist}(\cdot)$ gives the distance between a point and a target landmark, which is generally a curve or a region. The distance can take several forms; a simple one, widely used in previous works [Rabbani et al., 2022], akin to the Iterative Closest Point method, is to use the distance between the particle and the closest target landmark point, which is known to be differentiable [Fitzgibbon, 2003].

Since the MPM simulator is fully differentiable, gradients of the cost with respect to the initial particle velocities are computed by automatic differentiation across all

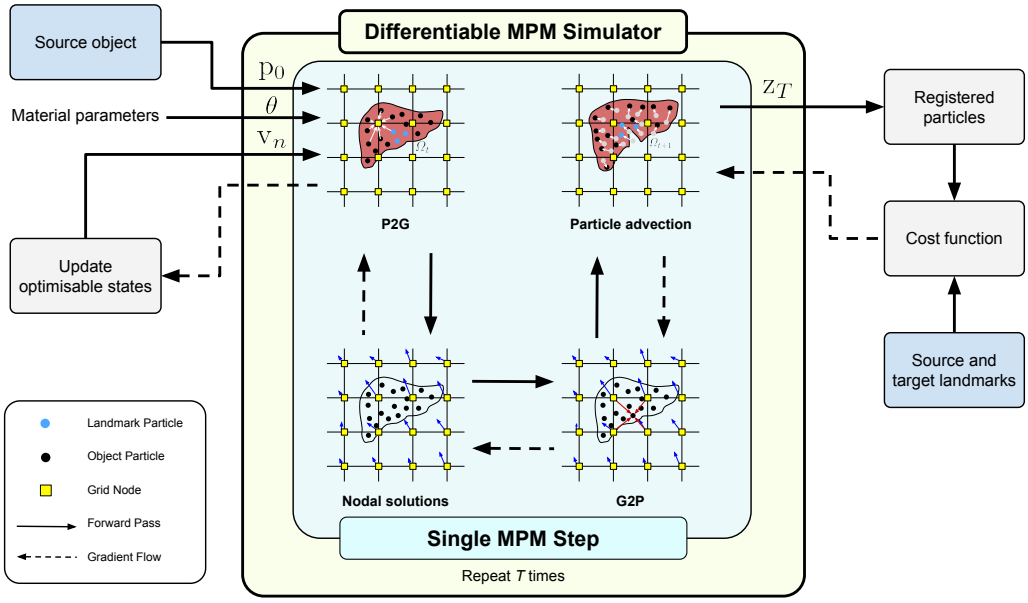


Figure 5.4: Overview of the proposed RODS-MPM framework. The source configuration is simulated using a differentiable MPM solver. A landmark-based cost is computed with respect to the target configuration, and gradients are backpropagated to optimise the initial particle velocities.

simulation steps. These gradients are used to iteratively update the initial state using gradient descent. Optimisation drives the simulated object toward alignment with the target configuration while ensuring that the resulting deformation remains physically consistent with the underlying material model. We refer to this differentiable physics-based registration method, illustrated in figure 5.4, as RODS-MPM.

5.4 Experimental Results

We evaluate the proposed RODS-MPM method in three progressively more challenging experimental settings. First, on a set of illustrative 2D examples, illustrating the capability of the method, second, we conduct *in-silico* experiments where realistic liver deformations with surgical incisions are simulated, enabling quantitative evaluation, and third, on a qualitative clinical experiment.

5.4.1 Illustrative Examples in 2D

We first demonstrate the behaviour of RODS-MPM on simple 2D geometric examples designed to illustrate both topology-preserving and topology-changing deformations. Figure 5.5 presents four scenarios in which a square domain, uniformly sampled with particles, undergoes deformations guided by different point and curve landmarks. The scenarios illustrate representative behaviours including a) large

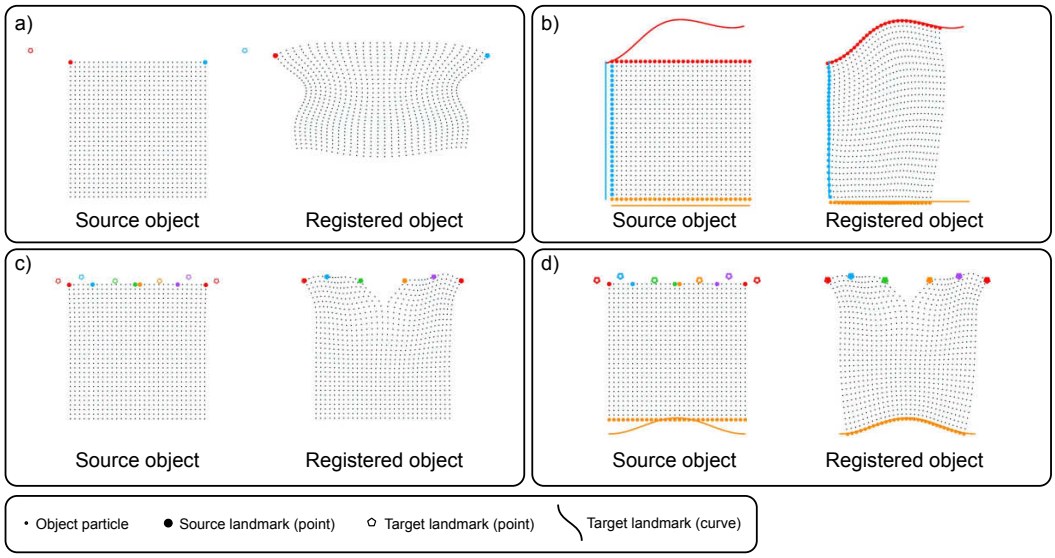


Figure 5.5: Illustrative examples showing the behaviour of RODS-MPM under different geometric constraints. The mesh-free particle representation allows topology changes to emerge naturally without explicit connectivity modification or remeshing.

elastic deformation, b) curve-landmark-driven registration, c) topological change (rupture), and d) combined large deformation and topology change. Depending on the task, the registration target is defined either through sparse point landmarks (a, b), through a distance function to a target curve (c), or through a combination of both (d). Within the RODS-MPM framework, topology changes arise naturally from the spatial configuration of these constraints.

It is worth noting that, in spite of landmark sparsity, the registered object naturally follows the deformation behaviour of hyperelastic materials. In particular, in figure 5.5(a), guided by only two point landmarks, RODS-MPM stretches the object horizontally while compressing its height, reproducing the characteristic response of elastic materials. In addition, landmark placements that induce separation behaviour, implicitly drive topology-changing deformations without requiring explicit modification of connectivity, as seen in figure 5.5(c) and (d).

5.4.2 In-Silico Experiments

To evaluate the proposed method under controlled conditions, we conduct a series of in-silico experiments on synthetically generated liver deformations with simulated surgical incisions. This setup enables precise control over deformation patterns and topology changes, while providing ground-truth correspondences for quantitative evaluation. We compare RODS-MPM against representative geometry-based and biomechanics-based registration methods.

Baseline methods. We compare RODS-MPM against four representative baselines. The first baseline is ARAP [Sorkine et al., 2007], a geometry-driven regularisation widely used in non-rigid registration. The second baseline is a standard Deformable Registration with Biomechanical Constraints Deformable Registration with Biomechanical Constraints (DRBC) based on [Collins et al., 2020]. This method formulates registration as the minimisation of a data fidelity term combined with a Neo-Hookean internal strain energy that acts as a regularisation term. The third baseline is Large Deformation Diffeomorphic Metric Mapping (LDDMM) [Beg et al., 2005]. The fourth baseline is the resection-aware registration method proposed in [François et al., 2021].

Dataset generation. To evaluate the behaviour of the proposed method under controlled conditions, we generate a synthetic dataset of liver deformations with simulated surgical incisions. Patient-specific 3D liver models reconstructed from preoperative imaging data [Rabbani et al., 2022] are used as the source anatomy and geometry, over which biomechanically realistic deformations with topology changes are simulated. Specifically, we consider the simulation of common surgical incisions in two clinically relevant scenarios: *(i)* a *non-anatomical resection*, typically performed for localised tumour removal with a 1 cm oncologic margin, and *(ii)* an *anatomical resection*, corresponding to the removal of an entire liver segment along anatomical boundaries.

The simulation pipeline consists of two stages. First, a surgical incision is created by subtracting a CAD model of a surgical blade from the volumetric liver mesh, introducing a geometric discontinuity in the mesh. Second, mechanical forces are applied to the two opposing surfaces of the incision using FEBio [Maas et al., 2012] to simulate tissue separation during surgery. FEBio (Finite Elements for Biomechanics) is a software package for finite element analysis and was specifically designed for applications in biomechanics and bioengineering.

The liver tissue is modelled as a nearly incompressible Neo-Hookean hyperelastic material with Young's modulus $E = 4.1$ kPa and Poisson's ratio $\nu = 0.49$, parameters commonly used for soft liver tissue simulation. To simulate tissue separation along the incision, equal and opposite forces are applied to the two sides of the cut surface, inducing progressive deformation and opening of the tissue during the simulation. The simulation is run for $T = 200$ time steps. The final configuration is used as the target geometry while the undeformed liver serves as the source model. The two sides of the incision boundary are annotated using polylines in both the source and target geometries. These polylines represent sparse surgical landmarks that guide the registration process. In a clinical setting, these landmarks can be automatically detected using a trained neural network, following an approach similar to that of [François et al., 2021].

Implementation details. RODS-MPM is implemented in DiffTaichi [Hu et al., 2019a] and executed on an NVIDIA A6000 GPU. The source volumetric liver model is uniformly sampled with a spacing of 0.6 mm to generate material particles, re-

sulting in approximately 6.8k particles. Particle velocities are initialised to zero. Material behaviour is modelled using a hyperelastic constitutive model parameterised by $\theta_1 = 4.1\,\text{kPa}$ and $\theta_2 = 0.49$. Source landmark particles are defined as the material particles closest to the annotated incision points in the source configuration. The registration target is represented as a polyline. The cost function measures the Euclidean distance between each source landmark particle and the target polyline. Gradients of this cost with respect to the initial particle velocities are computed through automatic differentiation across the differentiable MPM simulator. Optimisation is performed for 500 iterations using gradient descent with an initial learning rate of 25 and a decay factor of 0.5.

Results. Qualitative registration results are shown in figure 5.6, where the estimated deformation produced by each method is compared with the ground-truth deformation generated by the FEM simulation. The results clearly demonstrate that RODS-MPM successfully captures topology changes in both anatomical and non-anatomical resection scenarios. In contrast, ARAP and biomechanical FEM-based registration methods struggle to accommodate the incision-induced discontinuities. Because these methods rely on explicit mesh connectivity, they attempt to stretch the mesh to satisfy landmark constraints, resulting in unrealistic deformations near the incision. However, we observe that the deformation remains smooth far from the incision therefore classical models also work.

We also report the TRE measured at mesh nodes. Concretely, we measure the node-wise Euclidean distance between the registered model and the ground-truth model. Figure 5.7 presents spatial TRE maps for each method. Large errors are observed near the incision region for both baseline methods, whereas RODS-MPM maintains substantially lower errors due to its particle-based representation, which naturally allows separation of material. To analyse the spatial distribution of errors, figure 5.8 plots the mean and variance of the registration error as a function of the distance to the closest incision landmark. In the non-anatomical resection scenario, RODS-MPM achieves consistently lower errors both near and far from the incision region. In the anatomical resection scenario, the improvement is particularly stronger near the incision boundary, where topology changes occur.

Farther away from the incision region, all methods produce comparable results as the deformation becomes smoother and more consistent with classical elastic behaviour.

To evaluate the impact of landmark quantity and quality on the proposed method, we conducted additional experiments measuring the TRE under varying numbers of landmark points (20, 50, 250) and under different levels of Gaussian noise added to the landmark positions ($\sigma = 2.5, 5.0, 10.0\text{mm}$), as reported in table 5.1. The results demonstrate the robustness of the method to both sparse and noisy landmark configurations.

Sensitivity analysis. We evaluated the sensitivity of RODS-MPM to several key MPM parameters, including material properties, particle density, MPM grid reso-

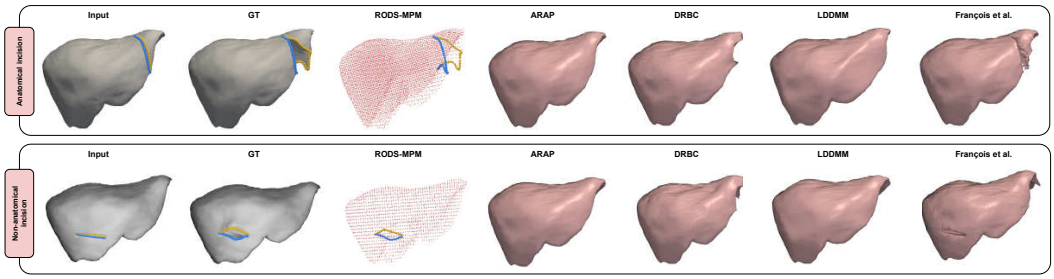


Figure 5.6: Qualitative comparison of deformable registration results under simulated surgical incisions. The source liver model is registered to the target landmarks for both anatomical and non-anatomical resections. Blue and yellow points represent landmarks located on the two sides of the incision boundary. RODS-MPM successfully reconstructs the tissue separation caused by the incision due to its particle-based representation. In contrast, ARAP, DRBC, and LDDMM are unable to accommodate topology changes and therefore produce unrealistic stretching near the incision region. Although the method of François et al. explicitly models topology changes, it still struggles to accurately recover the target geometry.

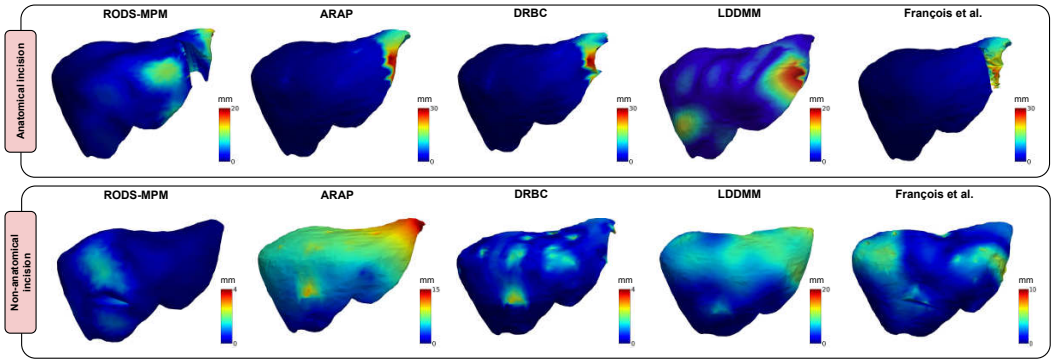


Figure 5.7: Spatial error maps of the registration results for the methods shown in figure 5.6. Errors are measured as the Euclidean distance between the registered nodes and the ground-truth model.

Table 5.1: TRE (mean ± standard deviation, mm).

	# landmark points			Gaussian noise σ (mm)		
	20	50	250	2.5	5.0	10.0
ARAP	3.12±2.54	2.05±2.43	1.95±2.08	2.82±2.92	3.04±3.12	3.41±3.48
DRBC	2.45±2.10	1.67±2.16	1.61±1.06	1.98±1.92	2.18±2.01	2.31±2.05
François et al.	2.31±2.18	1.80±1.86	1.72±1.14	1.93±1.87	2.08±1.99	2.24±1.91
LDDMM	4.50±3.23	3.46±2.86	2.83±2.26	3.73±3.07	4.24±3.52	4.54±3.89
Ours	**1.75±1.95**	**1.47±1.60**	**1.03±0.82**	**1.65±1.79**	**1.82±1.96**	**2.03±2.24**

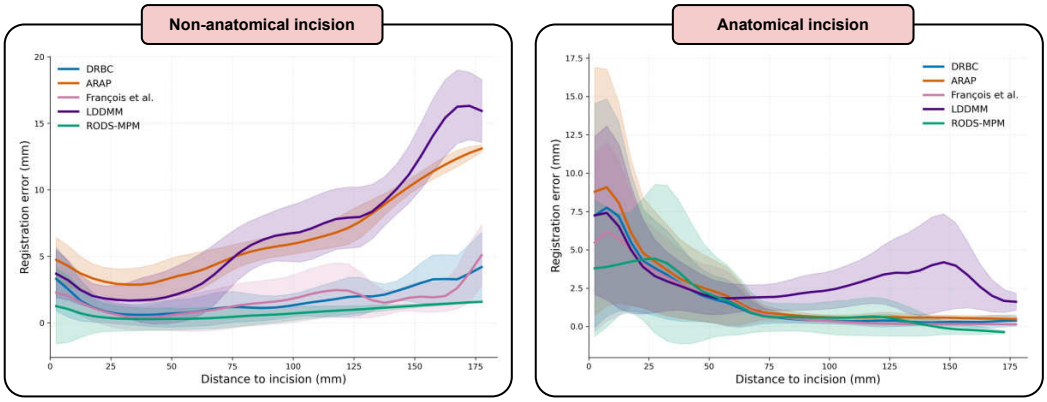


Figure 5.8: Quantitative evaluation of registration error as a function of the distance to the closest incision landmark. The plots show the mean registration error and its variance for each method.

lution, and simulation horizon (figure 5.9).

We assessed the sensitivity to material parameters by varying the Young's modulus around the reference value used in the experiments ($E = 4.1$ kPa), while keeping the Poisson ratio fixed at $\nu = 0.49$. Stiffness values ranging from $0.25\times$ to $10\times$ the reference modulus were evaluated. The results show stable performance for values approximately between $0.25\times$ and $3\times$, while larger stiffness values progressively degrade accuracy and increase variability, as overly stiff materials increasingly restrict the deformation required to match the target landmarks. A similar analysis performed for the Poisson ratio, while keeping the Young's modulus fixed, showed comparably stable behaviour. Overall, these results indicate that the proposed formulation does not critically depend on precise material calibration.

We further evaluated the impact of particle density. Increasing the number of particles generally improves accuracy, with the TRE progressively decreasing before reaching a plateau at higher densities. This behaviour indicates convergence with respect to the particle discretisation, where additional particles mainly provide a finer representation of the deformation field rather than substantially altering the recovered solution. Importantly, this behaviour remains independent of the particle resolution in terms of topology handling: increasing the particle density produces a finer and more detailed separation during topological changes, rather than enforcing continuity, since no constraint promotes reconnection between neighbouring particles once separation occurs.

We also varied the MPM grid resolution by changing n_{grid} from 30 to 180. TRE decreases as the grid is refined, with improvements becoming marginal beyond $n_{\text{grid}} \approx 100$, indicating convergence with respect to grid refinement. Although the variability moderately increases for some finer grids, the mean TRE remains stable, suggesting that excessive refinement mainly increases the sensitivity of the inverse optimisation without substantial accuracy gains. The resolution used in the

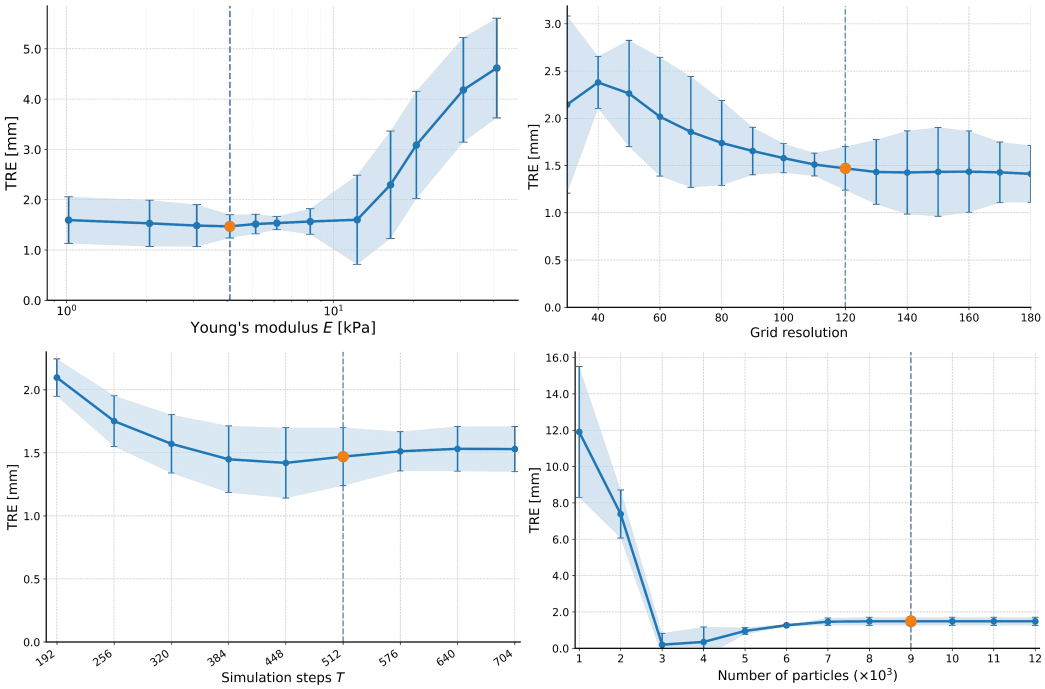


Figure 5.9: Sensitivity to MPM parameters. Orange circles indicate the experimental settings used in the experiments.

experiments ($n_{\text{grid}} = 120$) provides a reasonable trade-off between accuracy and computational cost.

Finally, increasing the simulation horizon progressively reduces the TRE before reaching a plateau, indicating that additional simulation steps initially improve the ability of the simulator to reproduce the target deformation, while beyond a certain point they provide only marginal gains.

5.4.3 Clinical Experiment

General points. We evaluate RODS-MPM retrospectively on stereoscopic images from a clinical liver surgery case, collected from a partner hospital under ethical clearance within an IRB approved protocol. The intraoperative data consist of a stereoscopic image sequence acquired during the procedure, while the preoperative 3D model of the liver is reconstructed from an MRI scan through a standard segmentation and meshing pipeline implemented in MITK. The stereoscopic camera is calibrated prior to the procedure using a checkerboard pattern and OpenCV.

The captured stereo images are rectified using the calibrated intrinsic and extrinsic camera parameters; depth is then recovered by computing disparity with the Foundation Stereo model [Wen et al., 2025], allowing the reconstruction of a dense intraoperative 3D point-cloud of the liver.

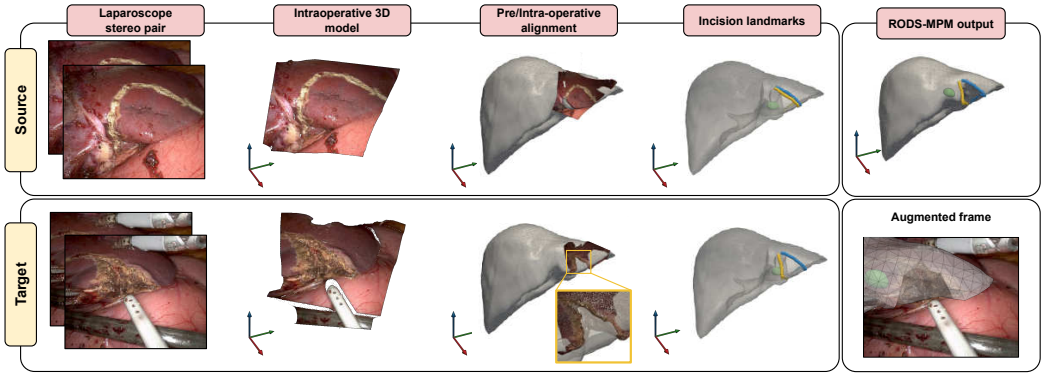


Figure 5.10: Registration of the preoperative liver model to intraoperative stereoscopic targets using RODS-MPM. The reconstructed intraoperative geometry and annotated anterior surface incision guide the deformation, and the resulting model is augmented onto the surgical image.

Data preparation. The task is to register the preoperative 3D model to the current stereo image pair. In a real-time surgical registration system, the preoperative 3D model is already registered to prior image pairs, whilst the system processes image pairs sequentially. In order to emulate these conditions, we select two stereo image pairs from the video. The first stereo image pair represents the past. It is taken before the incision occurs, but shortly after the surgeon marks the intended incision curve on the organ surface. The second stereo image pair represents the present. It is taken after the incision has been made. These choices were deliberately made to demonstrate RODS-MPM's ability to handle incisions. These images and the results are shown in figure 5.10.

In order to finalise the real working conditions, it remains to register the past image pair to the preoperative 3D model. For that, we manually align the latter to the reconstructed intraoperative surface manually, using visually identifiable anatomical landmarks. We also annotate the incision location in one of the two images and transfer it to the preoperative 3D model, completing the setup.

Registration with RODS-MPM We now have a preoperative 3D model comprising the incision boundary, serving as source landmarks for registration, and an image pair showing the incised organ, where the anterior surface incision boundary is marked, serving as target landmarks for registration. Importantly, the preoperative 3D model represents an uncut organ; in other words, there exists a strong deformation and topological changes between the preoperative 3D model and the organ shape as seen in the 'present' image pair.

The next step consists of computing the deformable registration. We run RODS-MPM whose inputs are the preoperative 3D model and the incision landmarks. The registered particles are then used to estimate the deformation field by simple barycentric coordinate interpolation, which is subsequently used to reg-

ister the internal structures, such as the tumour in our case. Since the output of RODS-MPM is a particle set, the resulting geometry is converted to a triangulated surface mesh using the Alpha Shapes algorithm [Edelsbrunner and Mücke, 1994] for display. Finally, the registered liver and the tumour it contains are projected back onto the stereo image and visualised as an augmented reality overlay on the target frame.

5.5 Discussion

In this chapter, we introduced RODS, a computational registration approach that formulates registration as an inverse problem over a differentiable physics simulator. By optimising the initial state of the simulator rather than directly estimating a deformation field, the method constrains the solution space to physically plausible trajectories. This provides an implicit regularisation driven by the underlying material model, reducing the reliance on hand-crafted deformation priors while improving realism under large and complex deformations.

We instantiated this framework using a differentiable MPM, resulting in the proposed RODS-MPM approach. Owing to its particle-based, mesh-free representation, MPM naturally accommodates topology changes such as cutting or separation, without requiring explicit remeshing or connectivity updates. This constitutes a key advantage over classical mesh-based registration methods, which inherently assume topology preservation and tend to produce unrealistic stretching or compression near incision boundaries. The proposed formulation therefore enables a unified treatment of both continuous deformations and topology-changing events within the same optimisation framework.

The experimental results support these observations. In illustrative 2D scenarios, RODS-MPM demonstrates the ability to reproduce physically consistent deformation patterns from sparse constraints, including cases involving rupture and separation. In controlled *in-silico* experiments, the method achieves significantly improved registration accuracy near incision regions compared to baselines, while maintaining comparable performance in regions where deformations remain smooth. The clinical experiment further highlights the practical relevance of the approach, showing that RODS-MPM can be integrated into a realistic surgical pipeline to register preoperative models to intraoperative observations in the presence of topology changes. While qualitative in nature, this experiment illustrates the potential of the method for AR guidance in surgery.

Despite these promising results, several limitations remain. First, the current formulation relies on predefined material parameters, which may not accurately reflect patient-specific tissue properties. Extending the framework to jointly identify material parameters alongside the initial state represents an important direction for future work. Second, the optimisation process requires multiple forward simulations, leading to higher computational cost compared to classical optimisation-

based or learning-based approaches. Improving efficiency, for instance through better initialisation strategies or reduced-order simulation models, would be necessary for real-time applications. Third, the method currently relies on sparse landmark constraints, whose accuracy and availability directly influence registration quality. Incorporating richer image-based constraints or dense correspondence cues could further improve robustness.

More broadly, the proposed approach highlights the potential of differentiable simulation as a paradigm for deformable registration. By shifting the problem from direct deformation estimation to optimisation over physically grounded simulation models, it becomes possible to handle complex behaviours that are difficult to encode explicitly, such as large deformations, contact, and topology changes. Future work will explore extensions toward more complex material models, and tighter integration with image-driven pipelines.

Chapter 6

Conclusion and Future Work

This thesis addressed several fundamental challenges in computer vision for MIS. In particular, it investigated complementary research directions including geometric structure recovery under challenging imaging conditions, stereo matching without camera calibration, robust correspondence estimation under strong perspective distortions, and deformable registration in the presence of complex tissue behaviour and topology changes. In the following, each contribution is briefly discussed together with its potential limitations and future research directions, and general directions are then outlined.

Robust depth estimation in MIS benchmark. To assess the reliability of MoSDE methods and overcome major limitations of existing datasets and benchmarks, we introduced RoDEM. Unlike prior approaches relying on sparse ground truth, RoDEM uses a dedicated depth sensing system capable of delivering dense depth maps at full frame rate. These depth maps are automatically aligned with laparoscopic images through a calibrated transformation, enabling acquisition under realistic surgical conditions. The dataset is publicly available, allowing the community to benchmark, train, and develop future methods. Despite its strengths, RoDEM still presents limitations. In particular, the use of ex-vivo data introduces a domain gap relative to in-vivo surgical conditions, while the depth sensing system remains affected by sensor-specific limitations that may influence the generated ground truth and evaluation results. RoDEM also highlighted the current limitations of metric MoSDE methods in MIS. Existing approaches are primarily trained on large-scale urban or outdoor datasets whose visual and geometric characteristics differ substantially from surgical scenes. Addressing this domain gap therefore represents an important future research direction, particularly given the potential value of accurate metric depth estimation for many surgical applications.

Camera augmentation training strategy. Concerning uncalibrated stereo matching, we introduced the CATS framework, which bridges the gap between idealised synthetic datasets and the imaging conditions encountered in MIS, where camera calibration is often uncertain or unavailable. By introducing controlled geometric perturbations of camera parameters during training, the proposed approach enables large-scale stereo matching models to adapt to surgical imaging systems. The CATS framework demonstrated that accurate and robust stereo matching can be achieved directly from uncalibrated image pairs while still maintaining compet-

itive performance in calibrated settings. A potential limitation might be that the CATS method was designed for near-rectified stereo laparoscopes and may be less effective for imaging systems exhibiting larger geometric deviations or stronger distortions. Regarding future work, the robustness demonstrated by the CATS framework suggests that large collections of publicly available stereo surgical datasets could be exploited to generate pseudo-ground-truth depth maps, even when calibration parameters are unavailable or unreliable. Another direction concerns the broader extension of the proposed CATS framework. Lastly, while this thesis focused primarily on da Vinci stereo laparoscopes, another future work could involve generalising the framework to a wider range of laparoscopic systems exhibiting different optical configurations, distortion models, baselines, and acquisition geometries.

Conformal perspective correction. CPC introduced an image warp designed to improve keypoint matching under strong perspective distortions on general surfaces. By transforming the image into a perspective-reduced representation through conformal flattening while controlling image resampling, CPC leads to more reliable matching. The experimental results support this motivation, showing consistent gains in difficult matching scenarios and positive effects on downstream tasks such as keyframe matching and 3D reconstruction. A potential limitation of the current framework is its suitability for surfaces exhibiting strong folds and severe self-occlusions. Extending CPC to better handle such complex structures represents an important direction for future work. Beyond methodological extensions, although this thesis established connections between perspective projection and conformality, a deeper mathematical analysis of the interaction between perspective projection and conformal transformations could provide stronger theoretical foundations for perspective correction methods.

Registration by optimisation over differentiable simulation. RODS introduced a deformable registration framework formulated as an inverse problem over a differentiable physics simulator. The method optimises the input variables of the simulator using a cost function defined on sparse observations, thereby recovering physically plausible deformations consistent with both the observations and the physical constraints of the simulator. Instantiated with a differentiable MPM, the resulting RODS-MPM framework additionally enables the modelling of topology-changing events. Nevertheless, the current formulation remains dependent on predefined material parameters, which may not fully reflect patient-specific tissue properties. Further future work could investigate material parameter identification. Another promising direction is the integration between differentiable simulation and differentiable rendering techniques. In particular, combining the MPM with neural rendering approaches such as 3DGS would allow dense photometric supervision directly from surgical images. Such formulations could constrain the deformation process using both geometric and appearance consistency, thereby improving registration accuracy and reducing ambiguity in underconstrained surgical settings. A

complementary direction is interactive surgical simulation and virtual cutting. An interesting extension of the proposed framework would involve simulating virtual surgical actions before actual tissue cutting occurs. For example, a surgeon could specify an incision trajectory or move a virtual surgical instrument over the organ surface without physically activating the cutting tool. The system would then simulate the predicted deformation and separation behaviour induced by the planned incision. Such predictive simulation tools could support surgical planning, training, and intraoperative decision-making by enabling surgeons to anticipate the consequences of surgical actions before execution.

Broader perspectives. Overall, the contributions of this thesis demonstrate that incorporating explicit geometric reasoning, camera modelling, and differentiable physical simulation substantially improves the robustness of vision-based surgical guidance systems. The proposed methods contribute toward bridging the gap between idealised computer vision assumptions and the challenging realities of intraoperative surgical environments. More broadly, computer vision for MIS is evolving toward increasingly multimodal, generative, and clinically integrated systems. In particular, multimodal surgical scene understanding is becoming an important research direction, combining endoscopic images with complementary sources of information such as robotic kinematics, force sensing, ultrasound, depth sensing, audio signals, and preoperative imaging. The increasing availability of robotic surgical platforms and multimodal sensing technologies offers new opportunities to improve robustness, contextual understanding, and intraoperative guidance. At the same time, recent advances in generative modelling open promising perspectives for simulation and synthetic data generation. Diffusion models, neural rendering, and video generation techniques could help address major limitations related to data scarcity, rare surgical events, and domain variability by enabling the generation of realistic and diverse surgical data under conditions that remain difficult to acquire clinically. Beyond these methodological developments, future research must also address the challenges associated with clinical translation. In particular, uncertainty estimation, explainability, reliable evaluation protocols, and safety guarantees will be essential for the safe and effective integration of computer vision systems into routine clinical practice.

Appendices

Appendix A: Guiding Breast Conservative Surgery by Augmented Reality

A.1 Context and Motivation

Breast cancer is one of the world's most prevalent malignancies, with approximately 12% of women encountering breast cancer [Fraser et al., 2016]. BCS is a common procedure to remove cancerous tumours. However, in cases where a tumour is not palpable, which represent over 50% of cases at diagnosis [Dua et al., 2011] but is detected in imaging modalities such as ultrasound, mammography or MRI, even expert surgeons have difficulties accurately localising it intraoperatively [Postma et al., 2011]. Moreover, with chemotherapy used prior to surgery, approximately 30% of the tumours will have a complete response and become extremely tiny or even vanish, making resection even more challenging [Spring et al., 2020].

Preoperative localisation of cancerous breast tumours involves different invasive modalities including Wire Guided Localisation (WGL), carbon tattooing, and, more recently, radioactive seed and magnetic seed localisation. WGL, also called needle localisation, is the most common method performed before BCS. This procedure is done by placing a fine thread-like wire close to the cancerous region or by targeting a biopsy clip marker deployed after the percutaneous biopsy at diagnosis. This way, the surgeon can follow the thread to reach the breast abnormality whose tissue must be extracted. This process is guided by the use of a mammogram or ultrasound. Therefore, this method usually involves two departments, radiology and surgical oncology, which complicates planning. From the patient's perspective, going through different procedures in two different settings is usually unpleasant as, even if inserting the guidance objects inside the breast is performed under local anaesthesia, it can be painful and is often very traumatic [Postma et al., 2012]. In addition, there is a small chance of wire displacement during confirmatory mammography or patient transferring [Postma et al., 2011]. Accurate localisation is essential to achieve complete resection and optimal cosmetic results. However, a significant pathological margin from the cancerous tissue may be obtained after the initial surgery. Thus, between 10% to 40% of the patients require at least one additional re-excision procedure to remove the remaining abnormal lesions [Jung et al., 2012, Fraser et al., 2016].

Recently, AR-based systems are introduced to enable surgeons to visualise the tumour location within the patient's breast using devices such as the Microsoft HoloLens [Perkins et al., 2017, Gouveia et al., 2021]. These approaches offer a new paradigm for real-time guidance during surgery, with the potential to improve localisation accuracy while reducing patient burden. However, existing systems are still research prototypes and share two main limitations caused by a) breast deformations and b) camera projection. We propose an AR system for BCS which uses preoperative MRI and an intraoperative RGB-D camera. We mitigate a) which mainly occurs because of gravity, by collecting a preoperative MRI in supine position. We mitigate b), which occurs because of variations in the relative breast to camera position, using a vertical projection method.

A.2 Proposed Method and Contributions

We propose an AR system for BCS which follows the same strategy as in MIS. The proposed system starts by reconstructing a digital twin as a preoperative 3D model for the breast and tumours. However, there is a strong difference: whilst MIS, whether traditional or robot-assisted, uses a camera and a screen, BCS is an open surgery, and uses neither of these. Therefore, we introduce both a camera and a screen in the OR, in order to capture the intraoperative surgical images and to display the augmented images for guidance. We have chosen an RGB-D camera, which provides both a regular colour image and a depth image in real time, as the technology is mature and low-cost. We exploit the depth image for 3D model registration. We have chosen a simple regular screen as display.

In contrast to existing AR systems for BCS, we propose the reconstruction of digital twins based on MRI acquired with patients in supine position. This approach mitigates the intense breast deformations caused by gravity. Further, unlike existing AR systems which render the tumours by direct virtual camera projection, we propose a vertical projection technique. Direct projection causes misguidance, as the resulting AR visualisation is then dependent on camera positioning. The proposed AR system for BCS has four steps, as illustrated in figure A.1.

Step 1). We begin with MRI data acquisition with the patient in supine position. This is in contrast with conventional MRI data acquisition for breast imaging, which is typically performed with the patient in prone position. The prone positioning is used because it allows the breast to hang away from the body, which helps in separating the breast tissue from the chest wall and provides a clearer image. This is particularly useful for detecting and characterising breast lesions and for surgical planning. However, as a result, the breast deforms intensely because of gravity. This huge deformation makes the registration phase of AR highly difficult. Recent studies [Fausto et al., 2020] assess the feasibility and image quality of breast MRI imaging performed in supine position compared to prone position. The study con-

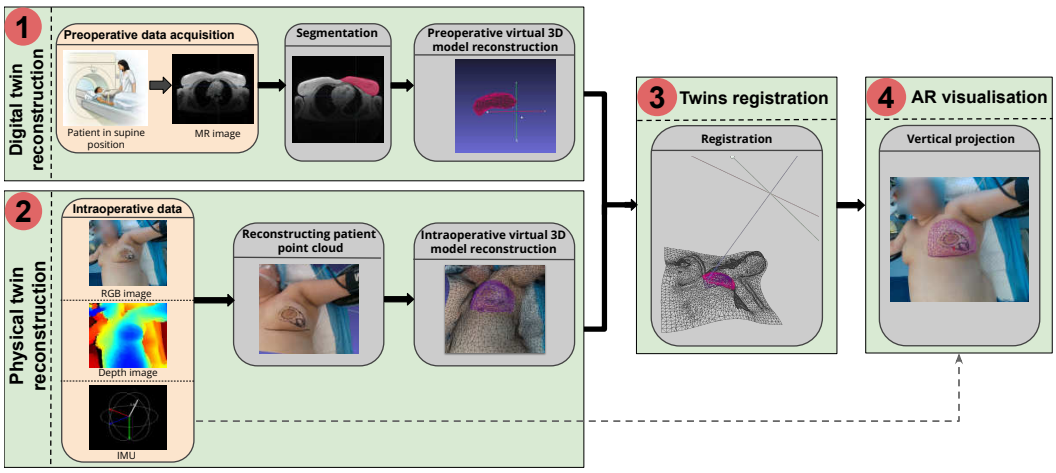


Figure A.1: Schematic of the proposed BCS guidance AR system. Inputs (orange), actions (gray), and main steps (green) are shown. The AR image is generated using the proposed vertical projection, with the gravity vector provided by the IMU.

cludes that there is no difference between supine and prone positioning in terms of image quality and number of lesions. However, a significant difference in the lesion extension and the breast shape can be observed comparing the two positions. Therefore, we use the preoperative MRI in supine position in order to mitigate intense breast deformations caused by gravity. The selected patients have a classical MRI acquisition in prone position and are then asked to lie down in supine position, with their arms alongside the body. An initial acquisition is performed without injection and then with injection of gadolinium, a commonly used contrast agent. With these MRI data, we reconstruct the preoperative 3D models of the desired structures. Concretely, we use MITK, to manually segment the breast and tumour on the MRI images, followed by reconstruction of their 3D models.

Step 2). We capture RGB-D images using an Intel RealSense camera at the beginning of surgery. We have used both the D415 and D435i models without noticing a difference in performance. This type of cameras provide conventional RGB images but also an image giving the depth of each pixel. Technically, using the camera parameters, we generate a point cloud of the scene, which is subsequently used to reconstruct the intraoperative model through Meshlab. More precisely, we first segment the region of interest from the obtained point cloud and reconstruct its surface using the screened Poisson surface reconstruction algorithm Kazhdan and Hoppe [2013].

Step 3). We register the 3D models obtained in steps 1) and 2). First, we perform an initial rough manual registration, focusing on key anatomical landmarks such as the nipple and aureola, as well as the breast silhouette. Second, we refine this registration using the Iterative Closest Point (ICP) algorithm. Technically, we use Meshlab for the first stage and Matlab for the second stage. This results in a rigid transformation which aligns the preoperative 3D model to the intraoperative model.

Step 4). We use the rigid transformation from step 3) to transfer the preoperative tumours 3D models to the surgical camera coordinates. With the tumours positioned correctly in their intraoperative location, we have two options for their visualisation in the AR system. The first option is direct projection, in which one projects the tumours towards the camera centre. This is the conventional practice in AR systems. However, a simple geometric reasoning as illustrated in figure A.2.a, shows that this strategy can lead to inconsistent renderings, due to variations in the relative positioning of the breast and camera. To address this, we examine a second approach for which we first project the tumour to a specific location on the breast surface and subsequently project to the camera, as illustrated in figure A.2.b. We perform the first projection following the gravity vector, for two main reasons. First, the surgeons commonly use a similar vertical projection, known as orthogonal localisation [Clough et al., 2003]. This method involves measuring the distances between the tumour and the nipple along the mediolateral and craniocaudal axes on both frontal and strict profile mammograms, and then transferring these measurements to the patient's breast. Following the same clinical strategy, we hypothesise that, given that the patient is positioned horizontally, the gravity vector is orthogonal to the breast, and thus leads to the same clinical orthogonal registration technique. Second, the gravity vector is typically readily available from the camera's Inertial Measurement Unit (IMU) sensor. In scenarios where the IMU data is unavailable, we fit a plane to the ground floor point cloud and use its normal as gravity vector. Finally, we project this region obtained on the breast surface to the camera to realise the AR overlay on the images captured by the RGB-D camera. Using this vertical projection, the virtually augmented tumour remains consistent.

A.3 Experimental Evaluation

Data. We evaluate the proposed method retrospectively in two BCS cases. All data were collected from hospital Centre Jean Perrin, Clermont-Ferrand, France, following the IRB approved protocol 00013468. The inclusion of only two patients was due to the rare availability of MRI in both the supine and prone positions. These specific patients were chosen by a radiologist who determined that a supine MRI sequence was necessary to assess the feasibility of breast-conserving surgery versus mastectomy. During intraoperative data collection, we ensured that the patient bed was positioned roughly parallel to the ground floor, which is standard for

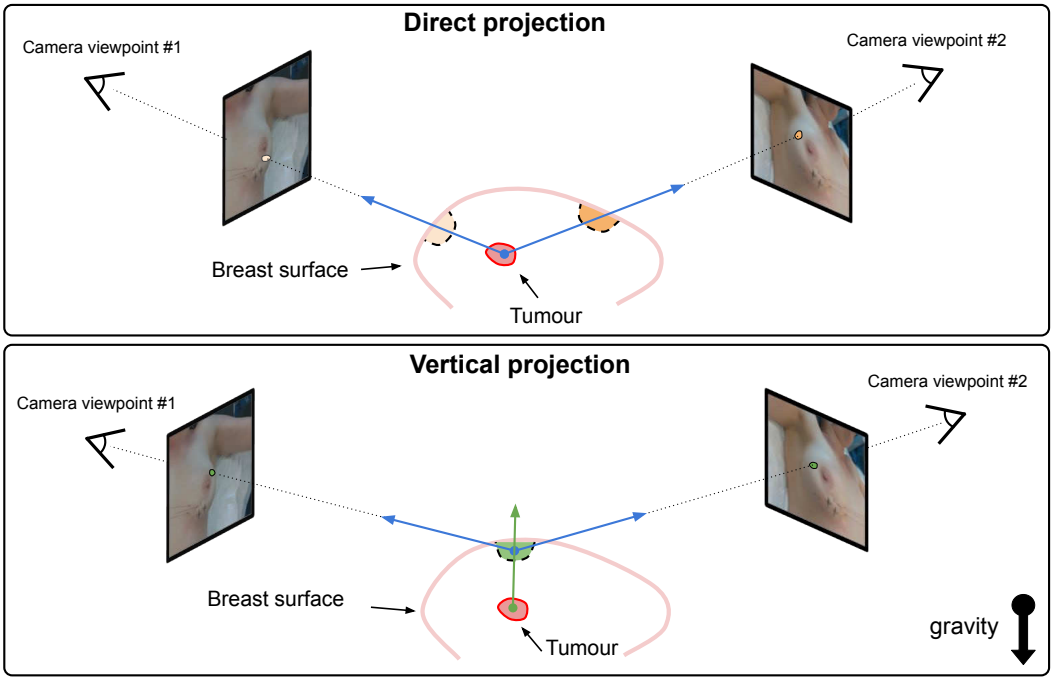


Figure A.2: Direct projection (top), where the tumour is projected directly to the camera, and vertical projection (bottom), where the tumour is first vertically projected to the breast surface along the gravity vector and then onto the camera. In vertical projection, the AR output remains consistent regardless of the camera positioning.

Table A.1: Residual registration error for the supine and prone position MRI.

	Patient #1		**Patient #2**	
	Supine position	**Prone position**	**Supine position**	**Prone position**
RMSE (mm)	**6.21**	34.71	**5.31**	28.14

BCS. We rotated the camera around the patient to examine the proposed system for different viewpoints. We selected three frames from the sequence representing extreme, middle and top views. We rendered AR visualisation using both direct projection and vertical projection methods, resulting in a total of 12 augmented images. For patient #1, the data was collected using a D415 camera, which lacks IMU data. Therefore, for this patient the gravity vector was estimated by computing the normal of the ground floor. For patient #2, the data was collected using a D435i camera which provides IMU information.

MRI acquisition. The impact of the MRI acquisition position on registration is illustrated by figure A.3. In prone positioning, deformable registration is required for accurate registration, whereas in supine positioning, a rigid transformation suffices for a fair alignment. We quantitatively evaluated the effectiveness of MRI

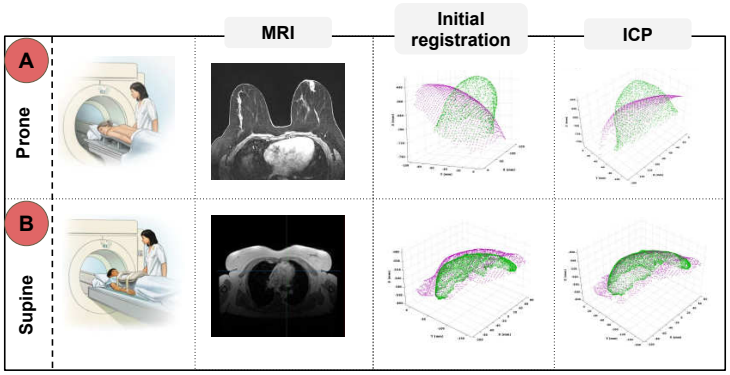


Figure A.3: Qualitative comparison of data acquisition and 3D model registration between different patient positioning. In A), the breast is intensely deformed due to gravity. Purple and green 3D points represent intraoperative and preoperative point clouds, respectively. The final rigid registration is significantly improved in supine position. The illustrated data pertain to patient #2.

acquisition in supine position by measuring the residual error obtained at the registration step. Specifically, we use the RMSE of ICP for both positioning. Concretely, we identified the closest point in the transformed point cloud for each point in the intraoperative model after registration. We then calculated the Euclidean distance between these points and computed the RMSE. Finally, we averaged the RMSE values obtained from three views. The results reported in table A.1 show significantly better alignment for supine position MRI.

Projection method. We performed quantitative evaluation of the proposed projection method by measuring the consistency of Euclidean distances between the tumours and anatomical landmarks in 3D. It is important to note that establishing a reliable ground truth for evaluating AR outputs in breast surgery has significant challenges. Although surgeons typically use the orthogonal localisation technique, however, this method has limitations. The transfer of coordinates from imagery is imprecise. Consequently, orthogonal registration lacks the sufficient reliability to serve as ground truth. Instead, we measured the 3D Euclidean distances between the centre of gravity of the nipple and the centre of gravity of the projected tumour on the breast surface, which should be constant. We report the standard deviation of these measurements as a metric representing the consistency of the projection methods in table A.2. The proposed vertical projection largely outperforms classical direct projection.

Expert evaluation. A surgeon evaluated the AR outputs as Very likely, Likely or Failure, based on the expected location of the rendered tumour. All the cases are shown in figure A.4. The results demonstrate an improvement in surgeon's overall satisfaction with the AR output when comparing the vertical to direct projection.

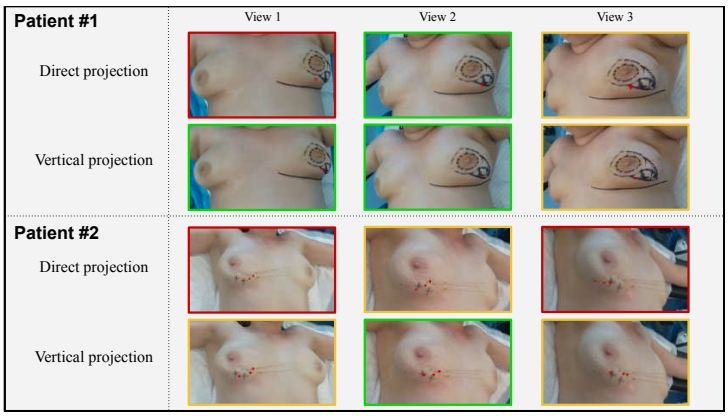


Figure A.4: Qualitative results of the proposed AR method. The augmented tumour is rendered in red. The image boundary colour shows the expert evaluation, with green for Very likely, yellow for Likely and red for Failure. The images were cropped to improve breast visualisation.

Table A.2: Standard deviation of measured 3D distances for both projection methods. Lower values indicate greater consistency in AR visualisation.

	Patient #1		**Patient #2**	
	Direct projection	**Vertical projection**	**Direct projection**	**Vertical projection**
Dist. std (mm)	23.15	**13.15**	31.43	**11.45**

For instance, for Patient #1 and View #1 the evaluation upgrades from Failure to Likely. Following the position of the rendered tumour in the sequence, we observe a more consistent AR output with the vertical projection for both patients.

A.4 Conclusion

We have proposed to use the principles of AR systems from MIS in open breast surgery. Implementing such a guidance system has many challenges, for which we have proposed solutions. Our system has shown promising results on two patients. The results reveal the importance of using a supine MRI and a vertical projection system. We now plan to 1) extend the validation of our system to a broader range of clinical cases, 2) incorporating deformable registration to account for remaining breast deformations, and 3) test our system intraoperatively.

Appendix B: Automatic Smoke Analysis in Minimally Invasive Surgery

This appendix summarises the work presented in [Sharifian et al., 2024]; readers are referred to the original publication for further details.

B.1 Problem Statement

Regular actions in MIS are to cut tissues, resect pathologies and handle bleeding. These procedures generally involve the use of electrocautery, relying on heated electrosurgical instruments such as monopolar, bipolar, or argon diathermy devices. Electrocautery generates surgical smoke composed primarily of water vapour along with organic and inorganic particles [Ulmer, 2008].

This smoke has two major negative consequences. First, it significantly degrades the visibility of the surgical scene, making it more difficult for the surgeon to operate and adversely affecting downstream computer vision tasks such as matching, reconstruction, and AR. Second, it contains potentially harmful substances for the surgical staff. These issues are mitigated by evacuating the smoke, either continuously using a smoke evacuation system inserted in an MIS port or by simply opening a port's valve. Manual smoke evacuation occurs upon request from the surgeon. We hypothesise that smoke management tasks may increase the surgeon's mental load. This hypothesis is supported by consistent feedback from the surgeons participating in this study, who have all suggested that surgical smoke management can impact their overall performance.

One possible solution is to perform systematic smoke evacuation as reported in [Takahashi et al., 2013], where smoke was evacuated automatically with a delay after electrosurgical device activation. The obvious drawback of systematic evacuation is that it activates even when not strictly required and may disrupt the intervention: it restores visibility but decreases the pressure of the pneumoperitoneum, the artificial insufflation required to create a workspace in the cavity. An alternative approach is to use continuous smoke evacuation by valveless trocars such as the Airseal insufflation system. Although this has been shown to have advantages [Balayssac et al., 2021], the method lacks efficiency since it performs evacuation constantly without considering the smoke situation, generating noise [Uppal et al., 2011] which

may have implications on sustainability of the operating room. It may also lead to potential postoperative complications [Weenink et al., 2020].

B.2 Proposed Method and Contributions

We propose a novel approach to smoke management, which is to automatically evaluate the need for smoke evacuation from the surgical image contents. Indeed, a central criterion to perform smoke evacuation is the lack of visibility caused by smoke. However, visually estimating the amount of smoke from an MIS image is challenging because of the wealth of other visual artefacts, including chromatic noise, spatial light variations, lens dirt and blur caused by camera or organ motion. Nevertheless, recent preliminary studies demonstrate the capability of neural networks for clear versus smoky image classification [Leibetseder et al., 2017a,b].

We therefore define our research question as the study of automatic smoke evacuation decisions from image analysis. To address this problem, we propose a methodology structured around three main contributions.

First, we construct a dataset of laparoscopy images with annotations collected from experts. Specifically, for each image, three types of annotations are collected: 1) the smoke level, which is a percentage representing the visual quantity of smoke present in the image, 2) the smoke evacuation confidence, which is a percentage representing the confidence in performing or not performing smoke evacuation, and 3) the smoke evacuation recommendation, which is a binary indicator representing the decision to evacuate or not evacuate smoke. We collected these annotations for a subset of the public datasets Smoke-Cholec80 [Leibetseder et al., 2017a] and LapGyn4-v1.2 [Leibetseder et al., 2018], and for a new gynaecology dataset collected from our hospital. For a part of the dataset, each image is annotated multiple times by multiple experts to analyse annotations consensus level.

Second, we perform a statistical analysis of inter-expert variability and investigate the relationships between the three types of annotations, namely, the smoke level, the smoke evacuation confidence and the smoke evacuation recommendation. Specifically, we study the extent to which the smoke evacuation confidence can be found from the smoke level and the extent to which the smoke evacuation recommendation can be found from the smoke level or from the smoke evacuation confidence.

Third, we leverage these relationships to propose image-based automatic prediction methods for three tasks, corresponding to the three types of annotations: 1) smoke quantification, 2) smoke evacuation confidence, and 3) smoke evacuation recommendation. To this end, we first train neural networks to solve these tasks in an end-to-end manner. We then introduce indirect methods, in which the output of the smoke quantification model is used to infer evacuation confidence and recommendation.

B.3 Results and Conclusion

Through performing multiple inter-expert analyses, we established a benchmark based on expert proficiency to determine the discrepancy between machine learning predictions and consensus annotations. We observe a reasonable inter-expert variability for smoke quantification and significantly higher variability for evacuation-related tasks. For smoke quantification, the expert error is 17.61 percentage points, while the neural network achieves a comparable error of 18.45 percentage points. For evacuation confidence, the best performance is obtained using the indirect predictor based on smoke quantification, achieving an error of 23.60 percentage points compared to an expert error of 27.35 percentage points. For evacuation recommendation, the expert accuracy is 76.78%, while the proposed method achieves 81.30%.

These results indicate that smoke quantification from still images can be reliably achieved by neural networks at a level comparable to human experts. In contrast, evacuation recommendation is inherently more challenging, even for experts, as reflected by the higher inter-expert variability. Nevertheless, we have shown that accurate smoke quantification can be leveraged to predict the need for evacuation on par with the experts. For more details on this appendix please refer to Sharifian et al. [2024].

Appendix C: Extended Results for RoDEM

C.1 Additional Qualitative Analysis

We examine the qualitative behaviour of the benchmarked methods under perturbation. Such analysis must explicitly account for the prediction-to-ground-truth alignment step described earlier, since this step compensates for scale and, when applicable, shift biases. In particular, the scale parameter captures an important portion of the global prediction bias. To illustrate these effects, we selected two representative perturbations: dark light variation, for which several methods exhibited unexpectedly strong robustness, and defocus blur, which proved particularly challenging. For each case, we computed error maps after scale-and-shift alignment, defined as the difference between aligned prediction and ground truth, expressed in millimetres. Figures C.1 and C.2 show the corresponding images together with the error maps for all benchmarked methods. In these visualisations, blue indicates that the predicted depth is larger than the ground truth, while red indicates that the prediction is shallower. Darker colours correspond to larger errors.

Several observations can be made. First, for both perturbations, the error generally increases with severity, although this increase is less pronounced for the most robust methods, particularly TRMDSV [Budd and Vercauteren, 2024a] and DA [Yang et al., 2024a]. Second, the liver surface tends to appear predominantly blue across many error maps, indicating that the methods often predict the liver as deeper than it actually is. This suggests a tendency to flatten the characteristic bumped geometry of the organ.

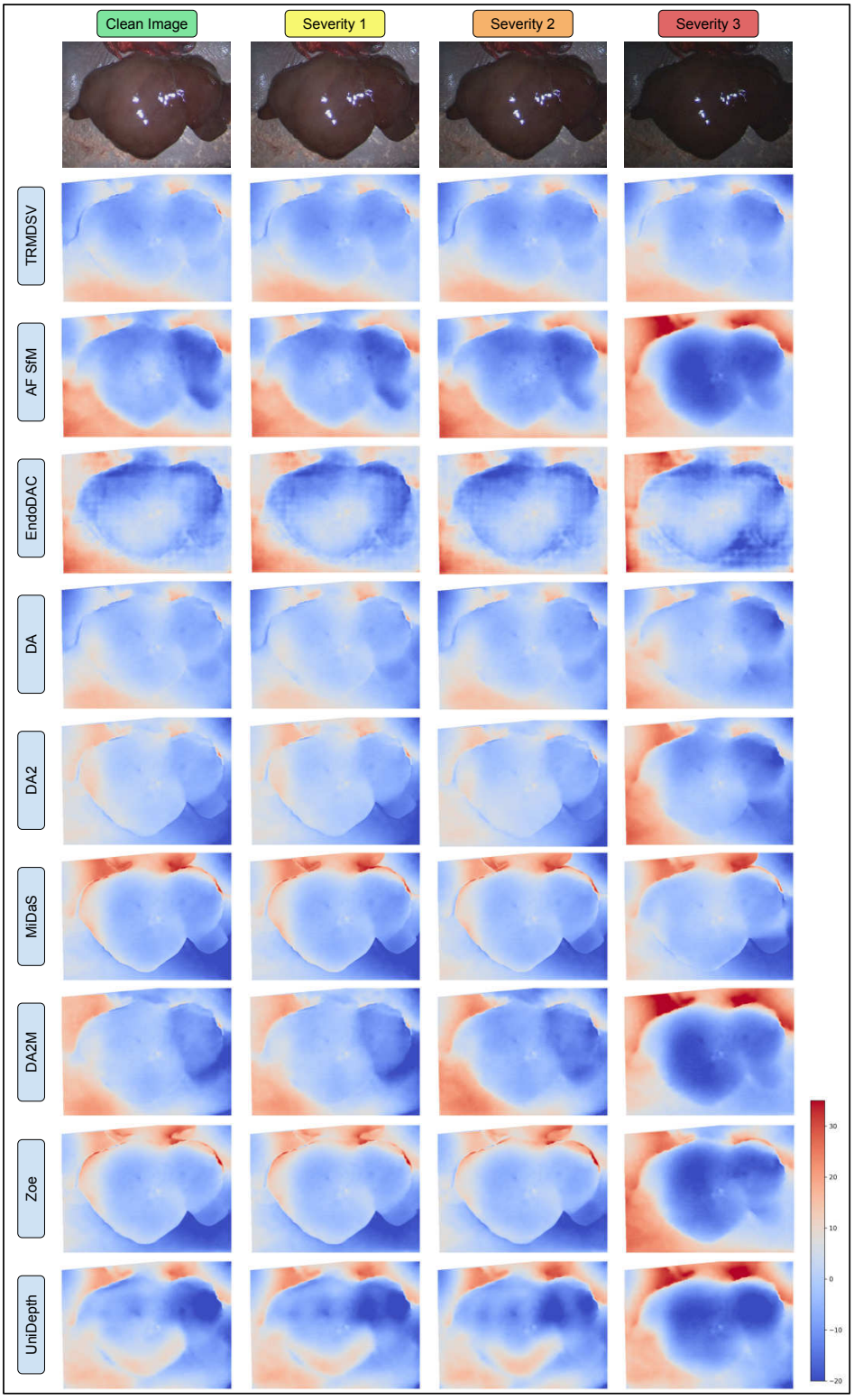


Figure C.1: Qualitative evaluation of the impact of the severity levels of dark light on the benchmarked methods.

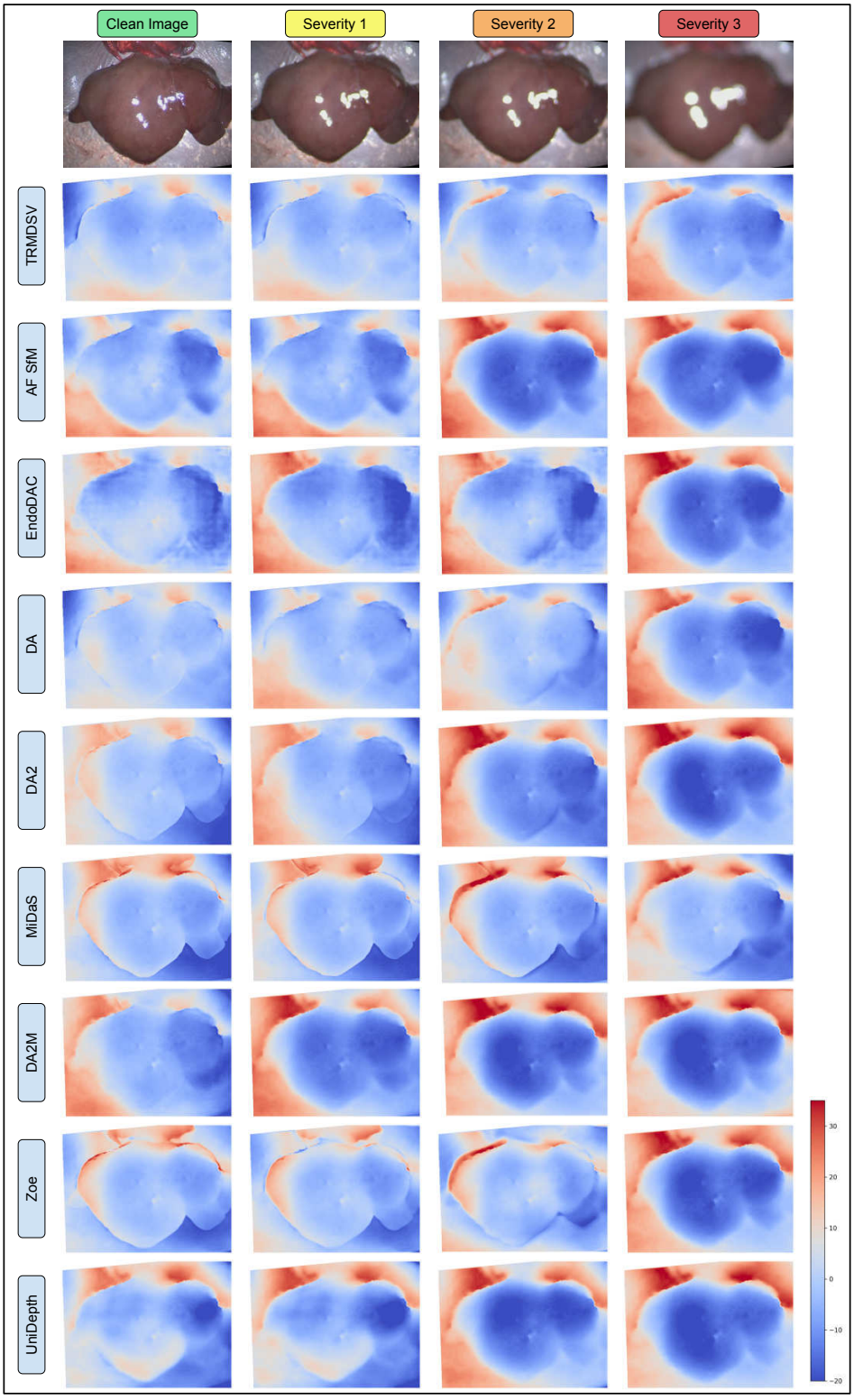


Figure C.2: Qualitative evaluation of the impact of the severity level of defocus blur on the benchmarked methods.

Scale and shift analysis. To further analyse this results, we examined the scale and shift parameters estimated during scale-and-shift alignment for the cases shown in figures C.1 and C.2. The corresponding values are reported in tables C.1 and C.2. When the shift remains approximately constant, the scale can be interpreted as a global depth bias: a larger scale indicates that the predicted disparity was, in a least-squares sense, globally smaller, and therefore that the corresponding depth prediction was farther away. For example, comparing the results for TRMDSV across dark light variation and defocus blur shows that the shift remains nearly unchanged, whereas the scale increases markedly under defocus. Since the organ, viewpoint, and overall scene remain the same, and only the amount of defocus changes, this suggests that TRMDSV tends to interpret defocus as increased depth.

Table C.1: Scales (first number) and shifts (second number) alignment estimations regarding the dark light perturbations presented in figure C.1.

	Clean Image	**Severity 1**	**Severity 2**	**Severity 3**
TRMDSV	0.249 - 3.360	0.239 - 3.372	0.237 - 3.350	0.247 - 3.317
AF-SfM	44.138 - 2.311	43.747 - 2.362	42.682 - 2.456	34.251 - 3.687
EndoDAC	51.143 - 1.378	50.700 - 1.408	54.875 - 1.236	61.500 - 0.914
DA	7.143e-02 - 2.916	7.353e-02 - 2.918	7.827e-02 - 2.902	7.675e-02 - 3.132
DA2	2.851e-02 - 3.147	2.759e-02 - 3.198	3.344e-02 - 3.066	3.127e-02 - 3.277
MiDaS	8.394e-04 - 3.400	8.568e-04 - 3.417	9.764e-04 - 3.371	1.005e-03 - 3.366
DA2M	1.721 - 1.569	1.966 - 1.392	2.022 - 1.352	0.793 - 4.345
Zoe	1.345 - 2.739	1.379 - 2.716	1.723 - 2.496	1.943 - 2.399
UniDepth	0.828 - 2.705	1.005 - 2.555	1.220 - 2.397	1.600 - 2.501

Table C.2: Scales (first number) and shifts (second number) alignment estimations regarding the defocus perturbations presented in figure C.2.

	Clean Image	**Severity 1**	**Severity 2**	**Severity 3**
TRMDSV	0.260 - 3.360	0.2548 - 3.377	0.303 - 3.412	0.499 - 3.466
AF-SfM	45.193 - 2.239	44.790 - 2.213	7.072 - 3.560	3.895 - 3.657
EndoDAC	41.677 - 1.863	30.197 - 2.377	38.821 - 1.944	12.302 - 3.242
DA	7.371e-02 - 2.929	7.133e-02 - 3.002	9.343e-02 - 2.923	5.936e-02 - 3.276
DA2	2.701e-02 - 3.238	2.408e-02 - 3.291	1.388e-02 - 3.572	2.196e-03 - 3.870
MiDaS	7.971e-04 - 3.460	7.118e-04 - 3.480	6.766e-04 - 3.510	1.050e-03 - 3.352
DA2M	1.457 - 1.885	0.517 - 3.017	0.060 - 3.940	0.195 - 4.311
Zoe	1.438 - 2.717	1.603 - 2.580	2.681 - 1.640	0.554 - 3.375
UniDepth	0.941 - 2.751	0.972 - 2.672	1.044 - 2.890	0.523 - 4.221

C.2 Additional Quantitative Analysis

We further report results for each perturbation type at different severity levels. For smoke, TRMDSV consistently achieves the best performance across severity levels, with DA typically ranking second. A similar trend is observed for blood, where TRMDSV remains the most accurate and most robust across all levels of severity. For lens dirtiness and defocus blur, performance degrades more substantially for all methods, although TRMDSV remains the strongest overall. These two perturbations are therefore confirmed as particularly challenging. For bright and dark light variations, the performance of the best methods remains relatively stable across severity levels, indicating that illumination changes are less detrimental than expected. Motion blur also has a comparatively limited effect, especially for the best-performing methods. Tables C.2.1 to C.16 summarise these severity-wise results for $\delta_0.25$ and Abs Rel under the scale-and-shift alignment setting.

C.2.1 Smoke

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.722	0.803	**0.964**	0.937	0.847	0.820	0.834	0.823	0.855
s1	0.713	0.808	**0.956**	0.951	0.873	0.807	0.856	0.830	0.827
s2	0.668	0.773	**0.958**	0.945	0.880	0.818	0.828	0.841	0.813
s3	0.604	0.737	**0.940**	0.915	0.834	0.816	0.723	0.817	0.707

Table C.3: $\delta_{0.25}$ metric with s&t-alignment for each method per severity

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.041	0.036	**0.022**	0.024	0.030	0.031	0.031	0.032	0.031
s1	0.041	0.035	0.022	**0.021**	0.028	0.032	0.029	0.031	0.034
s2	0.046	0.037	0.022	**0.021**	0.028	0.031	0.032	0.030	0.035
s3	0.053	0.040	**0.023**	0.024	0.032	0.032	0.041	0.033	0.043

Table C.4: Abs Rel metric with s&t-alignment for each method per severity

C.2.2 Blood

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.696	0.888	**0.977**	0.938	0.840	0.761	0.860	0.745	0.774
s1	0.546	0.841	**0.978**	0.922	0.779	0.725	0.811	0.711	0.734
s2	0.543	0.878	**0.973**	0.929	0.792	0.760	0.828	0.731	0.698
s3	0.431	0.858	**0.975**	0.936	0.802	0.727	0.872	0.696	0.694

Table C.5: $\delta_{0.25}$ metric with s&t-alignment for each method per severity

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.043	0.026	**0.017**	0.022	0.032	0.037	0.030	0.039	0.037
s1	0.056	0.030	**0.016**	0.024	0.036	0.039	0.034	0.041	0.041
s2	0.057	0.027	**0.017**	0.022	0.035	0.038	0.032	0.040	0.043
s3	0.066	0.029	**0.017**	0.024	0.034	0.039	0.030	0.042	0.044

Table C.6: Abs Rel metric with s&t-alignment for each method per severity

C.2.3 Dirty lens

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.715	0.790	**0.965**	0.931	0.857	0.825	0.875	0.813	0.813
s1	0.674	0.783	**0.938**	0.933	0.804	0.832	0.759	0.852	0.819
s2	0.688	0.764	**0.923**	0.848	0.783	0.826	0.734	0.865	0.776
s3	0.596	0.690	**0.813**	0.732	0.624	0.703	0.634	0.689	0.613

Table C.7: $\delta_{0.25}$ metric with s&t-alignment for each method per severity

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.042	0.036	**0.021**	0.024	0.029	0.032	0.029	0.033	0.034
s1	0.045	0.037	**0.024**	0.024	0.033	0.031	0.037	0.032	0.033
s2	0.044	0.037	**0.025**	0.029	0.036	0.033	0.040	0.030	0.038
s3	0.052	0.044	**0.032**	0.040	0.049	0.043	0.048	0.045	0.050

Table C.8: Abs Rel metric with s&t-alignment for each method per severity

C.2.4 Defocus blur

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.697	0.824	**0.975**	0.972	0.834	0.821	0.823	0.854	0.817
s1	0.695	0.832	**0.977**	0.937	0.838	0.824	0.713	0.860	0.783
s2	0.566	0.822	**0.936**	0.867	0.730	0.785	0.711	0.773	0.686
s3	0.513	0.661	**0.858**	0.799	0.661	0.616	0.582	0.569	0.547

Table C.9: $\delta_{0.25}$ metric with s&t-alignment for each method per severity

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.042	0.034	**0.019**	0.021	0.032	0.033	0.032	0.032	0.034
s1	0.044	0.033	**0.020**	0.024	0.032	0.033	0.040	0.032	0.038
s2	0.053	0.035	**0.024**	0.030	0.040	0.037	0.041	0.038	0.044
s3	0.059	0.043	**0.032**	0.034	0.046	0.051	0.053	0.053	0.059

Table C.10: Abs Rel metric with s&t-alignment for each method per severity

C.2.4.1 Light variations (bright)

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.706	0.787	**0.967**	0.935	0.853	0.816	0.833	0.831	0.850
s1	0.695	0.788	**0.961**	0.918	0.840	0.798	0.836	0.815	0.841
s2	0.738	0.786	**0.963**	0.914	0.850	0.833	0.823	0.834	0.844
s3	0.698	0.812	**0.956**	0.909	0.843	0.791	0.836	0.815	0.831

Table C.11: $\delta_{0.25}$ metric with s&t-alignment for each method per severity

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.042	0.036	**0.021**	0.024	0.030	0.032	0.031	0.032	0.032
s1	0.043	0.036	**0.021**	0.024	0.030	0.033	0.031	0.033	0.032
s2	0.040	0.036	**0.022**	0.025	0.029	0.031	0.032	0.031	0.033
s3	0.042	0.035	**0.021**	0.025	0.030	0.033	0.031	0.032	0.033

Table C.12: Abs Rel metric with s&t-alignment for each method per severity

C.2.4.2 Light variations (dark)

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.706	0.787	**0.967**	0.935	0.853	0.816	0.833	0.831	0.850
s1	0.631	0.758	**0.967**	0.915	0.827	0.817	0.815	0.815	0.838
s2	0.642	0.765	**0.961**	0.931	0.869	0.825	0.809	0.849	0.849
s3	0.560	0.696	**0.921**	0.871	0.822	0.810	0.716	0.822	0.796

Table C.13: $\delta_{0.25}$ metric with s&t-alignment for each method per severity

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.042	0.036	**0.021**	0.024	0.030	0.032	0.031	0.032	0.032
s1	0.047	0.038	**0.021**	0.025	0.031	0.032	0.032	0.032	0.033
s2	0.046	0.038	**0.022**	0.025	0.028	0.031	0.033	0.030	0.032
s3	0.053	0.043	**0.025**	0.029	0.033	0.032	0.040	0.032	0.034

Table C.14: Abs Rel metric with s&t-alignment for each method per severity

C.2.4.3 Motion blur

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.711	0.869	**0.960**	0.918	0.815	0.820	0.862	0.782	0.753
s1	0.664	0.886	**0.908**	0.837	0.751	0.697	0.761	0.642	0.655

Table C.15: $\delta_{0.25}$ metric with s&t-alignment for each method per severity

	AF-SfM	EndoDAC	TRMDSV	DA	DA2	MiDaS	DA2M	Zoe	UniDepth
s0	0.042	0.029	**0.020**	0.026	0.034	0.034	0.030	0.038	0.041
s1	0.047	0.028	**0.026**	0.031	0.040	0.044	0.037	0.049	0.048

Table C.16: Abs Rel metric with s&t-alignment for each method per severity

Bibliography

Nobuaki Akui, Satoshi Honma, Iwao Kanamori, Susumu Takahashi, Toyoharu Hanzawa, Takashi Fukaya, Hitoshi Karasawa, Tetsumaru Kubota, Toshihiko Hashiguchi, and Akihiko Mochida. Three-dimensional vision endoscope with position adjustment means for imaging device and visual field mask, 1996. US Patent 5,577,991. (Cited on page 47.)

Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J Davison. *KAZE features*, pages 214–227. Springer, 2012. (Cited on page 61.)

Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, and Sebastian Bodenstedt. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019. (Cited on page 46.)

Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, et al. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*, 2021a. (Cited on page 30.)

Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, et al. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*, 2021b. (Cited on pages 46, 48 and 50.)

Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2911–2918. IEEE, 2012. (Cited on pages 57 and 61.)

B. Braun. Minimally invasive surgery. `https://www.bbraun.com/en/products-and-solutions/therapies/minimally-invasive-surgery.html`, 2021. Accessed: 29 April 2026. (Cited on page 3.)

David Balayssac, Marie Selvy, Anthony Martelin, Caroline Giroudon, Delphine Cabelguenne, and Xavier Armoiry. Clinical and organizational impact of the airseal® insufflation system during laparoscopic surgery: a systematic review. *World journal of surgery*, 45(3):705–718, 2021. (Cited on page 123.)

Martin S Banks, Jenny CA Read, Robert S Allison, and Simon J Watt. Stereoscopy and the human visual system. *Motion imaging journal*, 121(4):24–43, 2012. (Cited on page 45.)

Joao Barreto, Jose Roquette, Peter Sturm, and Fernando Fonseca. Automatic camera calibration applied to medical endoscopy. In *BMVC 2009-20th British Machine*

Vision Conference, pages 1–10. The British Machine Vision Association (BMVA), 2009. (Cited on page 15.)

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18(153):1–43, 2018. (Cited on page 95.)

M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005. (Cited on page 101.)

J Lennart Berggren and Alexander Jones. *Ptolemy's Geography: An annotated translation of the theoretical chapters*. Princeton University Press, 2000. (Cited on page 63.)

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. (Cited on page 35.)

Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. (Cited on pages 18, 34 and 36.)

David E Blair. *Inversion theory and conformal mapping*, volume 9. American Mathematical Soc., 2000. (Cited on pages 66 and 67.)

Taylor L Bobrow, Mayank Golhar, Rohan Vijayan, Venkata S Akshintala, Juan R Garcia, and Nicholas J Durr. Colonoscopy 3d video dataset with paired depth from 2d-3d registration. *MedIA*, 2023. (Cited on page 29.)

Mario Botsch and Olga Sorkine. On linear variational surface deformation methods. *IEEE transactions on visualization and computer graphics*, 14(1):213–230, 2008. (Cited on page 89.)

Jeremiah U Brackbill, Douglas B Kothe, and Hans M Ruppel. Flip: a low-dissipation, particle-in-cell method for fluid flow. *Computer Physics Communications*, 48(1):25–38, 1988. (Cited on page 92.)

Robert S Breidenthal, Richard E Forkey, Jack Smith, and Brian E Volk. Stereoscopic endoscope with virtual reality viewing, 2000. US Patent 6,139,490. (Cited on page 47.)

Fred Brody and Nathan G Richards. Review of robotic versus conventional laparoscopic surgery. *Surgical endoscopy*, 28(5):1413–1424, 2014. (Cited on page 3.)

Duane C Brown. Close-range camera calibration. *Photogrammetric engineering*, 37 (8):855–866, 1971. (Cited on page 15.)

Matthew Brown and David G Lowe. Automatic panoramic image stitching using invariant features. *International journal of computer vision*, 74(1):59–73, 2007. (Cited on page 57.)

Charlie Budd and Tom Vercauteren. Transferring relative monocular depth to surgical vision with temporal consistency. In *MICCAI*, 2024a. (Cited on pages 79, 81, 82 and 127.)

Charlie Budd and Tom Vercauteren. Transferring relative monocular depth to surgical vision with temporal consistency. In *MICCAI*, 2024b. (Cited on pages 28 and 35.)

Kilian Chandelon, Rasoul Sharifian, Salomé Marchand, Abderrahmane Khaddad, Nicolas Bourdel, Nicolas Mottet, Jean-Christophe Bernhard, and Adrien Bartoli. Kidney tracking for live augmented reality in stereoscopic mini-invasive partial nephrectomy. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 11:1251–1260, 7 2023. ISSN 2168-1163. doi: 10.1080/21681163.2022.2157750. (Cited on pages 58 and 62.)

François Chaumette and Seth Hutchinson. Visual servo control. i. basic approaches. *IEEE robotics & automation magazine*, 13(4):82–90, 2006. (Cited on page 57.)

Xin-Zu Chen, Shao-Yong Wang, Yin-Su Wang, Zi-Han Jiang, Wei-Han Zhang, Kai Liu, Kun Yang, Xiao-Long Chen, Lin-Yong Zhao, Meng Qiu, et al. Comparisons of short-term and survival outcomes of laparoscopy-assisted versus open total gastrectomy for gastric cancer patients. *Oncotarget*, 8(32):52366, 2017. (Cited on page 4.)

Priya Bhave Chittawar, Sebastian Franik, Annefloor W Pouwer, and Cindy Farquhar. Minimally invasive surgical techniques versus open myomectomy for uterine fibroids. *Cochrane Database of Systematic Reviews*, 2014. (Cited on page 4.)

Krishna B. Clough, Dominique Heitz, and Rémy J. Salmon. Chirurgie locorégionale des cancers du sein. In *Encyclopédie médico-chirurgicale: Techniques chirurgicales – Gynécologie*, pages 15–31. Elsevier SAS, Paris, 2003. (Cited on page 118.)

Toby Collins, Daniel Pizarro, Simone Gasparini, Nicolas Bourdel, Pauline Chauvet, Michel Canis, Lilian Calvet, and Adrien Bartoli. Augmented reality guided laparoscopic surgery of the uterus. *IEEE Transactions on Medical Imaging*, 40(1):371–380, 2020. (Cited on pages 8, 85, 89 and 101.)

Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. (Cited on page 85.)

Claude Couinaud. *Le foie; études anatomiques et chirurgicales*. Masson, 1957. (Cited on page 87.)

A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *International Journal of Computer Vision*, 40:123–148, 2000. ISSN 09205691. doi: 10.1023/A:1026598000963. (Cited on page 62.)

Beilei Cui, Mobarakol Islam, Long Bai, An Wang, and Hongliang Ren. Endodac: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In *MICCAI*, 2024. (Cited on pages 28, 35, 76, 79, 81 and 82.)

Alban De Vaucorbeil, Vinh Phu Nguyen, Sina Sinaie, and Jian Ying Wu. Material point method after 25 years: Theory, implementation, and applications. *Advances in applied mechanics*, 53:185–398, 2020. (Cited on page 92.)

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 337–33712. IEEE, 6 2018. ISBN 978-1-5386-6100-0. doi: 10.1109/CVPRW.2018.00060. (Cited on pages 61 and 75.)

Sascha Miles Dua, Richard J. Gray, and Mohammed Keshtgar. Strategies for localisation of impalpable breast lesions. *The Breast*, 20:246–253, 6 2011. ISSN 09609776. doi: 10.1016/j.breast.2011.01.007. (Cited on page 115.)

Herbert Edelsbrunner and Ernst P Mücke. Three-dimensional alpha shapes. *ACM Transactions On Graphics (TOG)*, 13(1):43–72, 1994. (Cited on page 107.)

PJ Eddie Edwards, Dimitris Psychogyios, Stefanie Speidel, Lena Maier-Hein, and Danail Stoyanov. Serv-ct: A disparity dataset from cone-beam ct for validation of endoscopic 3d reconstruction. *MedIA*, 2022. (Cited on pages 29 and 30.)

Yamid Espinel, Lilian Calvet, Karim Botros, Emmanuel Buc, Christophe Tilmant, and Adrien Bartoli. Using multiple images and contours for deformable 3d-2d registration of a preoperative ct in laparoscopic liver surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 657–666. Springer, 2021. (Cited on page 85.)

Alfonso Fausto, Annarita Fanizzi, Luca Volterrani, Francesco Giuseppe Mazzei, Claudio Calabrese, Donato Casella, Marco Marcasciano, Raffaella Massafra, Daniele La Forgia, and Maria Antonietta Mazzei. Feasibility, image quality and clinical evaluation of contrast-enhanced breast mri performed in a supine position compared to the standard prone position. *Cancers*, 12:2364, 8 2020. ISSN 2072-6694. doi: 10.3390/cancers12092364. (Cited on page 116.)

Bernd Fischer and Jan Modersitzki. Curvature based image registration. *Journal of Mathematical Imaging and Vision*, 18(1):81–85, 2003. (Cited on page 88.)

Bernd Fischer and Jan Modersitzki. Ill-posed medicine—an introduction to image registration. *Inverse problems*, 24(3):034008, 2008. (Cited on page 88.)

Andrew W Fitzgibbon. Robust registration of 2d and 3d point sets. *Image and vision computing*, 21(13-14):1145–1153, 2003. (Cited on page 98.)

Harley Flanders. Liouville's theorem on conformal mapping. *Journal of Mathematics and Mechanics*, 15(1):157–161, 1966. (Cited on page 68.)

Michael S Floater and Kai Hormann. Surface parameterization: a tutorial and survey. *Advances in multiresolution for geometric modelling*, pages 157–186, 2005. (Cited on page 65.)

Tom François, Lilian Calvet, Callyane Sève-d'Erceville, Nicolas Bourdel, and Adrien Bartoli. Image-based incision detection for topological intraoperative 3d model update in augmented reality assisted laparoscopic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 647–656. Springer, 2021. (Cited on pages 8, 90 and 101.)

Victoria J Fraser, Katelin B Nickel, Ida K Fox, Julie A Margenthaler, and Margaret A Olsen. The epidemiology and outcomes of breast cancer surgery. *Transactions of the American Clinical and Climatological Association*, 127:46–58, 2016. ISSN 0065-7778. (Cited on page 115.)

GeigerandAndreas, RoserandMartin, and UrtasunandRaquel. Efficient large-scale stereo matching. In *ACCV*, 2010. (Cited on page 30.)

Robert A Gingold and Joseph J Monaghan. Smoothed particle hydrodynamics: theory and application to non-spherical stars. *Monthly notices of the royal astronomical society*, 181(3):375–389, 1977. (Cited on page 91.)

Joshua Gluckman and Shree K Nayar. Rectifying transformations that minimize resampling effects. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. (Cited on pages 68 and 82.)

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. (Cited on page 28.)

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. (Cited on page 62.)

Pedro F. Gouveia, Joana Costa, Pedro Morgado, Ronald Kates, David Pinto, Carlos Mavioso, João Anacleto, Marta Martinho, Daniel Simões Lopes, Arlindo R. Ferreira, Vasileios Vavourakis, Myrianthi Hadjicharalambous, Marco A. Silva, Nickolas Papanikolaou, Celeste Alves, Fatima Cardoso, and Maria João Cardoso.

Breast cancer surgery with augmented reality. *The Breast*, 56:14–17, 4 2021. ISSN 09609776. doi: 10.1016/j.breast.2021.01.004. (Cited on page 116.)

Oscar G Grasa, Ernesto Bernal, Santiago Casado, Ismael Gil, and JMM Montiel. Visual slam for handheld monocular endoscope. *IEEE transactions on medical imaging*, 33(1):135–146, 2013. (Cited on pages 26 and 57.)

Rostam Affendi Hamzah and Haidi Ibrahim. Literature survey on stereo vision disparity map algorithms. *Journal of Sensors*, 2016(1):8742920, 2016. (Cited on page 16.)

Francis H Harlow. The particle-in-cell method for numerical solution of problems in fluid dynamics. Technical report, Los Alamos Scientific Lab., N. Mex., 1962. (Cited on page 92.)

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. (Cited on pages 20, 21 and 57.)

Michel Hayoz, Christopher Hahne, Mathias Gallardo, Daniel Candinas, Thomas Kurmann, Maximilian Allan, and Raphael Sznitman. Learning how to robustly estimate camera pose in endoscopic videos. *IJCARS*, 18(7):1185–1192, 2023. (Cited on page 46.)

Eric Heiden, David Millard, Erwin Coumans, Yizhou Sheng, and Gaurav S Sukhatme. Neuralsim: Augmenting differentiable simulators with neural networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9474–9481. IEEE, 2021. (Cited on pages 95 and 96.)

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019. (Cited on page 37.)

Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *TPAMI*, 47(1):7–42, 2007. (Cited on page 16.)

Kai Hormann, Bruno Lévy, and Alla Sheffer. Mesh parameterization: Theory and practice. *ACM SIGGRAPH 2007 Courses, Association for Computing Machinery*, pages 1–115, 2007. (Cited on page 65.)

Berthold KP Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. *MIT AI lab*, 1970. (Cited on page 27.)

Yuanming Hu, Tzu-Mao Li, Luke Anderson, Jonathan Ragan-Kelley, and Frédo Durand. Taichi: a language for high-performance computation on spatially sparse data structures. *ACM Transactions on Graphics (TOG)*, 38(6):1–16, 2019a. (Cited on pages 95 and 101.)

Yuanming Hu, Jiancheng Liu, Andrew Spielberg, Joshua B Tenenbaum, William T Freeman, Jiajun Wu, Daniela Rus, and Wojciech Matusik. Chainqueen: A real-time differentiable physical simulator for soft robotics. In *2019 International conference on robotics and automation (ICRA)*, pages 6265–6271. IEEE, 2019b. (Cited on page 95.)

Yuanming Hu, Luke Anderson, Tzu-Mao Li, Qi Sun, Nathan Carr, Jonathan Ragan-Kelley, and Frédo Durand. DiffTaichi: Differentiable programming for physical simulation. *International Conference on Learning Representations (ICLR)*, 2020. (Cited on page 96.)

Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araujo. Omniglue: Generalizable feature matching with foundation model guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19865–19875, 2024. (Cited on page 61.)

Woohyun Jung, Eunyoung Kang, Sun Mi Kim, Dongwon Kim, Yoonsun Hwang, Young Sun, Cha Kyong Yom, and Sung-Won Kim. Factors associated with re-excision after breast-conserving surgery for early-stage breast cancer. *Journal of Breast Cancer*, 15:412, 2012. ISSN 1738-6756. doi: 10.4048/jbc.2012.15.4.412. (Cited on page 115.)

Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. (Cited on page 117.)

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. (Cited on page 13.)

Jae-Hak Kim, Adrien Bartoli, Toby Collins, and Richard Hartley. Tracking by detection for interactive image augmentation in laparoscopy. In *International Workshop on Biomedical Image Registration*, pages 246–255. Springer, 2012. (Cited on page 62.)

Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, and Wei Tsang Ooi. Robodepth: Robust out-of-distribution depth estimation under corruptions. *NeurIPS*, 2024. (Cited on pages 31 and 37.)

Seiichi Koshizuka and Yoshiaki Oka. Moving-particle semi-implicit method for fragmentation of incompressible fluid. *Nuclear science and engineering*, 123(3):421–434, 1996. (Cited on page 91.)

Gregory Kramida. Resolving the vergence-accommodation conflict in head-mounted displays. *IEEE transactions on visualization and computer graphics*, 22(7):1912–1931, 2015. (Cited on page 47.)

Mathieu Labrunie. *Contributions to the Automation of Preoperative 3D Model to 2D Image Registration in Mini-Invasive Liver Surgery: Primitive Detection, Pose Estimation, Registration and Anatomical Modelling*. PhD thesis, Université Clermont Auvergne, 2024. URL http://dx.doi.org/10.70675/d2f7a3ccz0020z41edzb473z77254303a4f4. (Cited on page 89.)

Mathieu Labrunie, Daniel Pizarro, Christophe Tilmant, and Adrien Bartoli. Automatic 3d/2d deformable registration in minimally invasive liver resection using a mesh recovery network. In *Medical imaging with deep learning*, 2023. (Cited on page 8.)

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE, 2006. (Cited on page 57.)

Andreas Leibetseder, Manfred Jürgen Primus, Stefan Petscharnig, and Klaus Schoeffmann. Image-based smoke detection in laparoscopic videos. In *International Workshop on Computer-Assisted and Robotic Endoscopy*, pages 70–87. Springer, 2017a. (Cited on page 124.)

Andreas Leibetseder, Manfred Jürgen Primus, Stefan Petscharnig, and Klaus Schoeffmann. Real-time image-based smoke detection in endoscopic videos. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 296–304, 2017b. (Cited on page 124.)

Andreas Leibetseder, Stefan Petscharnig, Manfred Jürgen Primus, Sabrina Kletz, Bernd Münzer, Klaus Schoeffmann, and Jörg Keckstein. Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology. In *Proceedings of the 9th ACM multimedia systems conference*, pages 357–362, 2018. (Cited on page 124.)

Ling Li, Xiaojian Li, Bo Ouyang, Hangjie Mo, Hongliang Ren, and Shanlin Yang. Three-dimensional collision avoidance method for robot-assisted minimally invasive surgery. *Cyborg and Bionic Systems*, 4:0042, 2023. (Cited on pages 26 and 43.)

Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. In *3DV*, pages 218–227, 2021. (Cited on pages 42 and 49.)

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 11 2004. ISSN 0920-5691. doi: 10.1023/B:VISI.0000029664.99615.94. URL http://link.springer.com/10.1023/B:VISI.0000029664.99615.94. (Cited on page 60.)

Michael Lutter. A differentiable newton–euler algorithm for real-world robotics. In *Inductive Biases in Machine Learning for Robotics and Control*, pages 9–34. Springer, 2023. (Cited on page 95.)

Bruno Lévy, Sylvain Petitjean, Nicolas Ray, and Jérome Maillot. Least squares conformal maps for automatic texture atlas generation. *ACM Transactions on Graphics*, 21:362–371, 7 2002. ISSN 0730-0301. doi: 10.1145/566654.566590. URL https://dl.acm.org/doi/10.1145/566654.566590. (Cited on pages 65 and 70.)

Steve A Maas, Benjamin J Ellis, Gerard A Ateshian, and Jeffrey A Weiss. Febio: finite elements for biomechanics. *Journal of Biomechanical Engineering*, 2012. (Cited on page 101.)

Miles Macklin. Warp: A high-performance python framework for gpu simulation and graphics. https://github.com/nvidia/warp, March 2022. NVIDIA GPU Technology Conference (GTC). (Cited on pages 95 and 96.)

Lena Maier-Hein, Peter Mountney, Adrien Bartoli, Haytham Elhawary, D Elson, Anja Groch, Andreas Kolb, Marcos Rodrigues, J Sorger, Stefanie Speidel, et al. Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. *Medical image analysis*, 17(8):974–996, 2013. (Cited on pages 7, 15 and 57.)

Shivali Malhotra, Osama Halabi, Sarada Prasad Dakua, Jhasketan Padhan, Santu Paul, and Waseem Palliyali. Augmented reality in surgical navigation: a review of evaluation and validation metrics. *Applied Sciences*, 13(3):1629, 2023. (Cited on page 26.)

Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016. (Cited on page 48.)

K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65:43–72, 11 2005. ISSN 0920-5691. doi: 10.1007/s11263-005-3848-x. (Cited on page 60.)

Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60:63–86, 10 2004. ISSN 0920-5691. doi: 10.1023/B:VISI.0000027790.02288.f2. (Cited on pages 60 and 61.)

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. (Cited on page 13.)

Raymond David Mindlin. Second gradient of strain and surface-tension in linear elasticity. *International journal of solids and structures*, 1(4):417–438, 1965. (Cited on page 88.)

Dmytro Mishkin, Jiri Matas, Michal Perdoch, and Karel Lenc. Wxbs: Wide baseline stereo generalizations. *arXiv preprint arXiv:1504.06603*, 2015. (Cited on page 61.)

Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In *Proceedings of the European conference on computer vision (ECCV)*, pages 284–300, 2018. (Cited on page 61.)

Jan Modersitzki. *Numerical methods for image registration*. OUP Oxford, 2003. (Cited on pages 85 and 88.)

Richard Modrzejewski, Toby Collins, Barbara Seeliger, Adrien Bartoli, Alexandre Hostettler, and Jacques Marescaux. An in vivo porcine dataset and evaluation methodology to measure soft-body laparoscopic liver registration accuracy with an extended algorithm that handles collisions. *International journal of computer assisted radiology and surgery*, 14(7):1237–1245, 2019. (Cited on page 30.)

Richard Modrzejewski, Toby Collins, Alexandre Hostettler, Jacques Marescaux, and Adrien Bartoli. Light modelling and calibration in laparoscopy. *International journal of computer assisted radiology and surgery*, 15(5):859–866, 2020. (Cited on page 7.)

Joe J Monaghan. Smoothed particle hydrodynamics. *Reports on progress in physics*, 68(8):1703–1759, 2005. (Cited on page 91.)

Jean-Michel Morel and Guoshen Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2:438–469, 1 2009. ISSN 1936-4954. doi: 10.1137/080732730. (Cited on page 61.)

Peter Mountney, Danail Stoyanov, and Guang-Zhong Yang. Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Process. Mag.*, 2010. (Cited on page 30.)

Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31 (5):1147–1163, 2015. (Cited on page 57.)

J. K. Murthy, Miles Macklin, Florian Golemo, Vikram Voleti, Ludovic Petrini, Martin Weiss, Brennan Considine, Jonathan Parent-Lévesque, Junjie Xie, Stefano Ermon, et al. gradSim: Differentiable simulation for system identification and visuomotor control. In *International Conference on Learning Representations (ICLR)*, 2021. (Cited on page 96.)

Sanjeev Narasimhan, Mehmet Kerem Turkcan, Mattia Ballo, Sarah Choksi, Filippo Filicori, and Zoran Kostic. Monocular 3d tooltip tracking in robotic surgery—building a multi-stage pipeline. *Electronics*, 14(10):2075, 2025. (Cited on pages 26 and 43.)

Rhys Newbury, Jack Collins, Kerry He, Jiahe Pan, Ingmar Posner, David Howard, and Akansel Cosgun. A review of differentiable simulators. *IEEE Access*, 12: 97581–97604, 2024a. (Cited on page 95.)

Rhys Newbury, Jack Collins, et al. A review of differentiable simulators. *IEEE Access*, 2024b. Early Access. (Cited on pages 95 and 96.)

Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. (Cited on page 85.)

Vinh Phu Nguyen, Alban De Vaucorbeil, and Stephane Bordas. The material point method. *Cham: Springer International Publishing*, 2023. (Cited on page 92.)

Ilana Nisky, Felix Huang, Amit Milstein, Carla M Pugh, Ferdinando A Mussa-Ivaldi, and Amir Karniel. Perception of stiffness in laparoscopy–the fulcrum effect. *Studies in health technology and informatics*, 173:313, 2012. (Cited on page 4.)

Somayeh Norouzi-Ghazbi and Farrokh Janabi-Sharifi. Dynamic modeling and system identification of internally actuated, small-sized continuum robots. *Mechanism and Machine Theory*, 154:104043, 2020. (Cited on page 95.)

Raymond William Ogden. Large deformation isotropic elasticity–on the correlation of theory and experiment for incompressible rubberlike solids. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 326(1567): 565–584, 1972. (Cited on page 89.)

Victor Okunrintemi, Faiz Gani, and Timothy M Pawlik. National trends in postoperative outcomes and cost comparing minimally invasive versus open liver and pancreatic surgery. *Journal of Gastrointestinal Surgery*, 20(11):1836–1843, 2016. (Cited on page 4.)

Christoph J Paulus, Nazim Haouchine, Seong-Ho Kong, Renato Vianna Soares, David Cazier, and Stéphane Cotin. Handling topological changes during elastic registration: application to augmented reality in laparoscopic surgery. *International journal of computer assisted radiology and surgery*, 12(3):461–470, 2017. (Cited on pages 8 and 90.)

Alex Paul Pentland. A new sense for depth of field. *IEEE transactions on pattern analysis and machine intelligence*, pages 523–531, 1987. (Cited on page 28.)

Veronica Penza, Andrea S Ciullo, Sara Moccia, Leonardo S Mattos, and Elena De Momi. Endoabs dataset: Endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms. *Int. J. Med. Robot.*, 2018. (Cited on page 30.)

Stephanie L. Perkins, Michael A. Lin, Subashini Srinivasan, Amanda J. Wheeler, Brian A. Hargreaves, and Bruce L. Daniel. A mixed-reality system for breast surgical planning. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pages 269–274. IEEE, 10 2017. ISBN 978-0-7695-6327-5. doi: 10.1109/ISMAR-Adjunct.2017.92. (Cited on page 116.)

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, 2024. (Cited on page 35.)

Julien Pilet, Vincent Lepetit, and Pascal Fua. Fast non-rigid surface detection, registration and realistic augmentation. *IJCV*, 76(2):109–122, 2008. (Cited on page 85.)

E. L. Postma, H. M. Verkooijen, S. van Esser, M. G. Hobbelink, G. P. van der Schelling, R. Koelemij, A. J. Witkamp, C. Contant, P. J. van Diest, S. M. Willems, I. H. M. Borel Rinkes, M. A. A. J. van den Bosch, W. P. Mali, and R. van Hillegersberg. Efficacy of 'radioguided occult lesion localisation' (roll) versus 'wire-guided localisation' (wgl) in breast conserving surgery for non-palpable breast cancer: a randomised controlled multicentre trial. *Breast Cancer Research and Treatment*, 136:469–478, 11 2012. ISSN 0167-6806. doi: 10.1007/s10549-012-2225-z. (Cited on page 115.)

Emily L Postma, Arjen J Witkamp, Maurice AAJ van den Bosch, Helena M Verkooijen, and Richard van Hillegersberg. Localization of nonpalpable breast lesions. *Expert Review of Anticancer Therapy*, 11:1295–1302, 8 2011. ISSN 1473-7140. doi: 10.1586/era.11.116. (Cited on page 115.)

Philip Pratt, Danail Stoyanov, Marco Visentini-Scarzanella, and Guang-Zhong Yang. Dynamic guidance for robotic surgery using image-constrained biomechanical models. In *MICCAI*, 2010. (Cited on page 30.)

Chandrakanth Jayachandran Preetha, Jonathan Kloss, Fabian Siegfried Wehrtmann, Lalith Sharan, Carolyn Fan, Beat Peter Müller-Stich, Felix Nickel, and Sandy Engelhardt. Towards augmented reality-based suturing in monocular laparoscopic training. In *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 11315, pages 232–238. SPIE, 2020. (Cited on pages 26 and 43.)

Philip Pritchett and Andrew Zisserman. Wide baseline stereo matching. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 754–760. IEEE, 1998. (Cited on page 57.)

James Pritts, Ondrej Chum, and Jiri Matas. Detection, rectification and segmentation of coplanar repeated patterns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980, 2014. (Cited on page 62.)

Dimitrios Psychogyios, Evangelos Mazomenos, Francisco Vasconcelos, and Danail Stoyanov. MSDESIS: Multitask stereo disparity estimation and surgical instrument segmentation. *TMI*, 41(11):3218–3230, 2022. (Cited on pages 18, 42 and 50.)

Claudius Ptolemy. The geography, translated by el stevenson, 1991. (Cited on page 63.)

Navid Rabbani, Lilian Calvet, Yamid Espinel, Bertrand Le Roy, Mathieu Ribeiro, Emmanuel Buc, and Adrien Bartoli. A methodology and clinical dataset with ground-truth to evaluate registration accuracy quantitatively in computer-assisted laparoscopic liver resection. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 10(4):441–450, 2022. (Cited on pages 98 and 101.)

Prem Rashid, Jeremy Goad, Monish Aron, Troy Gianduzzo, and Inderbir S Gill. Laparoscopic partial nephrectomy: integration of an advanced laparoscopic technique. *ANZ journal of surgery*, 78(6):471–475, 2008. (Cited on page 4.)

David Recasens, José Lamarca, José M Fácil, JMM Montiel, and Javier Civera. Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE RA-L*, 2021. (Cited on page 30.)

Timothy Rengers and Susanne Warner. Surgery for colorectal liver metastases: anatomic and non-anatomic approach. *Surgery*, 174(1):119–122, 2023. (Cited on page 87.)

Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3674–3683, 2020. (Cited on page 75.)

Bernhard Riemann. *Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse*. PhD thesis, Universität Göttingen,, 1851. (Cited on page 65.)

Ronald S Rivlin. Large elastic deformations of isotropic materials iv. further developments of the general theory. *Philosophical transactions of the royal society of London. Series A, Mathematical and physical sciences*, 241(835):379–397, 1948. (Cited on page 89.)

TN Robinson and GV Stiegmann. Minimally invasive surgery. *Endoscopy*, 36(01): 48–51, 2004. (Cited on page 2.)

Mariano Rodríguez, Gabriele Facciolo, Rafael Grompone von Gioi, Pablo Musé, and Julie Delon. Robust estimation of local affine maps and its applications to image matching. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1342–1351, 2020. (Cited on page 61.)

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. (Cited on page 72.)

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. (Cited on page 61.)

Daniel Rueckert, Alejandro F Frangi, and Julia A Schnabel. Automatic construction of 3D statistical deformation models of the brain using nonrigid registration. *IEEE transactions on medical imaging*, 22(8):1014–1025, 2003. (Cited on page 85.)

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. (Cited on pages 61 and 75.)

Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1):7–42, 2002. (Cited on page 42.)

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. (Cited on page 77.)

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. (Cited on page 77.)

Shuwei Shao, Zhongcai Pei, Weihai Chen, Wentao Zhu, Xingming Wu, Dianmin Sun, and Baochang Zhang. Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. *MedIA*, 2022. (Cited on pages 28, 35, 79 and 82.)

R. Sharifian, H.M. Abrão, S. Madad-Zadeh, C. Seve, P. Chauvet, N. Bourdel, M. Canis, and A. Bartoli. Automatic smoke analysis in minimally invasive surgery by image-based machine learning. *Journal of Surgical Research*, 296, 2024. ISSN 10958673. doi: 10.1016/j.jss.2024.01.008. (Cited on pages 123 and 125.)

Rasoul Sharifian, Navid Rabbani, Yongcong Zhang, and Adrien Bartoli. Camera augmentation: enabling uncalibrated stereo matching of minimally invasive surgery images by training from the wealth of public synthetic image datasets. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–9, 2026. (Cited on page 7.)

Manjunath Siddaiah-Subramanya, Kor Woi Tiang, and Masimba Nyandowe. A new era of minimally invasive surgery: progress and development of major technical innovations in general surgery over the last decade. *The Surgery Journal*, 3(04): e163–e166, 2017. (Cited on page 4.)

Eftychios Sifakis and Jernej Barbic. FEM simulation of 3D deformable solids: a practitioner's guide to theory, discretisation and model reduction. In *Acm siggraph 2012 courses*, pages 1–50. ACM, 2012. (Cited on pages 88 and 89.)

Simeon J Simoff, Michael H Böhlen, and Arturas Mazeika. Visual data mining: An introduction and overview. In *Visual data mining: theory, techniques and tools for visual analytics*, pages 1–12. Springer, 2008. (Cited on page 57.)

Changkyu Song and Abdeslam Boularias. Learning to slide unknown objects with differentiable physics simulations. *arXiv preprint arXiv:2005.05456*, 2020. (Cited on page 95.)

Olga Sorkine, Marc Alexa, et al. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116, 2007. (Cited on pages 71, 89 and 101.)

Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE transactions on medical imaging*, 32(7):1153–1190, 2013. (Cited on page 88.)

Laura M. Spring, Geoffrey Fell, Andrea Arfe, Chandni Sharma, Rachel Greenup, Kerry L. Reynolds, Barbara L. Smith, Brian Alexander, Beverly Moy, Steven J. Isakoff, Giovanni Parmigiani, Lorenzo Trippa, and Aditya Bardia. Pathologic complete response after neoadjuvant chemotherapy and impact on breast cancer recurrence and survival: A comprehensive meta-analysis. *Clinical Cancer Research*, 26:2838–2848, 6 2020. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-19-3492. (Cited on page 115.)

Danail Stoyanov, George P Mylonas, Fani Deligianni, Ara Darzi, and Guang Zhong Yang. Soft-tissue motion tracking and structure estimation for robotic assisted mis procedures. In *MICCAI*, 2005. (Cited on page 30.)

Danail Stoyanov, Marco Visentini Scarzanella, Philip Pratt, and Guang-Zhong Yang. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In *MICCAI*, 2010. (Cited on page 30.)

SM Strasberg, J Belghiti, P-A Clavien, E Gadzijev, JO Garden, W-Y Lau, M Makuuchi, and RW Strong. The brisbane 2000 terminology of liver anatomy and resections. *Hpb*, 2(3):333–339, 2000. (Cited on page 87.)

Deborah Sulsky, Zhen Chen, and Howard L Schreyer. A particle method for history-dependent materials. *Computer methods in applied mechanics and engineering*, 118(1-2):179–196, 1994. (Cited on page 92.)

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. (Cited on pages 57 and 75.)

Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022. (Cited on page 28.)

Hidekazu Takahashi, Makoto Yamasaki, Masashi Hirota, Yasuaki Miyazaki, Jeong Ho Moon, Yoshihito Souma, Masaki Mori, Yuichiro Doki, and Kiyokazu Nakajima. Automatic smoke evacuation in laparoscopic surgery: a simplified method for objective evaluation. *Surgical endoscopy*, 27(8):2980–2987, 2013. (Cited on page 123.)

Shaharyar Ahmed Khan Tareen and Zahra Saleem. A comparative analysis of sift, surf, kaze, akaze, orb, and brisk. In *2018 International conference on computing, mathematics and engineering technologies (iCoMET)*, pages 1–10. IEEE, 2018. (Cited on page 61.)

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020. (Cited on page 50.)

Nita Thiruchelvam, Ser Yee Lee, and Adrian Kah Heng Chiow. Patient and port positioning in laparoscopic liver resections. *Hepatoma Research*, 7:N–A, 2021. (Cited on page 4.)

Carl Toft, Daniyar Turmukhambetov, Torsten Sattler, Fredrik Kahl, and Gabriel J Brostow. Single-image depth prediction makes feature matching easier. In *European conference on computer vision*, pages 473–492. Springer, 2020. (Cited on pages 61 and 62.)

Giorgos Tolias and Hervé Jégou. Visual query expansion with or without geometry: Refining local descriptors by feature aggregation. *Pattern Recognition*, 47:3466–3476, 10 2014. ISSN 00313203. doi: 10.1016/j.patcog.2014.04.007. (Cited on page 61.)

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE TMI*, 2016. (Cited on page 29.)

Brenda C Ulmer. The hazards of surgical smoke. *AORN journal*, 87(4):721–738, 2008. (Cited on page 123.)

Shitanshu Uppal, Michael Frumovitz, Pedro Escobar, and Pedro T Ramirez. Laparoendoscopic single-site surgery in gynecology: review of literature and available technology. *Journal of Minimally Invasive Gynecology*, 18(1):12–23, 2011. (Cited on page 123.)

An Wang, Haochen Yin, Beilei Cui, Mengya Xu, and Hongliang Ren. Benchmarking robustness of endoscopic depth estimation with synthetically corrupted data. In *SASHIMI*, 2024a. (Cited on pages 29 and 31.)

Di Wang, Hua Liu, and Xiang Cheng. A miniature binocular endoscope with local feature matching and stereo matching for 3d measurement and 3d reconstruction. *Sensors*, 18(7):2243, 2018. (Cited on page 26.)

Shuzhe Wang, Juho kannala, Marc Pollefeys, and Daniel Barath. Guiding local feature matching with surface curvature. In *ICCV*, 2023. (Cited on page 61.)

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *CVPR*, pages 20697–20709, 2024b. (Cited on page 50.)

RP Weenink, M Kloosterman, R Hompes, PJ Zondervan, HP Beerlage, PJ Tanis, and RA van Hulst. The airseal® insufflation device can entrain room air during routine operation. *Techniques in coloproctology*, 24(10):1077–1082, 2020. (Cited on page 124.)

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-shot stereo matching. In *CVPR*, pages 5249–5260, 2025. (Cited on pages 16, 18, 42, 48, 49 and 105.)

Keenon Werling, Dalton Omens, Jeongseok Lee, Ioannis Exarchos, and C Karen Liu. Fast and feature-complete differentiable physics for articulated rigid bodies with contact. *arXiv preprint arXiv:2103.16021*, 2021. (Cited on page 96.)

Jian Wu, Zhiming Cui, Victor S. Sheng, Pengpeng Zhao, Dongliang Su, and Shengrong Gong. A comparative study of sift and its variants. *Measurement Science Review*, 13:122–131, 6 2013. ISSN 1335-8871. doi: 10.2478/msr-2013-0021. (Cited on page 61.)

Renkai Wu, Changyu He, Pengchen Liang, Yinghao Liu, Yiqi Huang, Weiping Liu, Biao Shu, Panlong Xu, and Qing Chang. MCF-SMSIS: multi-tasking with complementary functions for stereo matching and surgical instrument segmentation. *Computers in Biology and Medicine*, 179:108923, 2024. (Cited on pages 42 and 50.)

Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Junda Cheng, Chunyuan Liao, and Xin Yang. IGEV++: Iterative multi-range geometry encoding volumes for stereo matching. *TPAMI*, 47(8):7108–7122, 2025. (Cited on pages 42 and 49.)

Mengya Xu, Ziqi Guo, An Wang, Long Bai, and Hongliang Ren. A review of 3d reconstruction techniques for deformable tissues in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 157–167. Springer, 2024. (Cited on page 26.)

Tetsuzo Yamaguchi, Masahiko Nakamoto, Yoshinobu Sato, Kozo Konishi, Makoto Hashizume, Nobuhiko Sugano, Hideki Yoshikawa, and Shinichi Tamura. Development of a camera model and calibration procedure for oblique-viewing endoscopes. *Computer Aided Surgery*, 9(5):203–214, 2004. (Cited on page 15.)

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024a. (Cited on pages 34 and 127.)

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024b. (Cited on pages 34 and 35.)

Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European conference on computer vision*, pages 467–483. Springer, 2016. (Cited on page 61.)

Bernhard Zeisl, Kevin Köser, and Marc Pollefeys. Viewpoint invariant matching via developable surfaces. In *European Conference on Computer Vision*, pages 62–71. Springer, 2012. (Cited on page 62.)

Wenyi Zhao, Catherine Mohr, and Simon DiMaio. Stereo imaging system with automatic disparity adjustment for displaying close range objects, 2019. US Patent 10,178,368. (Cited on page 47.)

Zhengming Zhou and Qiulei Dong. Learning occlusion-aware coarse-to-fine depth map for self-supervised monocular depth estimation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6386–6395, 2022. (Cited on page 28.)

Jiayuan Zhu, Yunli Qi, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024. (Cited on page 34.)

Yongning Zhu and Robert Bridson. Animating sand as a fluid. *ACM Transactions on Graphics (TOG)*, 24(3):965–972, 2005. (Cited on page 91.)