Enhancing Fluoroscopy-Guided Procedures with Neural Network-Based Deformable Organ Registration

François Lecomte

Thèse de doctorat

École doctorale Mathématiques, sciences de l'information et de l'ingénieur | ED 269
Université de Strasbourg

Université de Strasbourg

Enhancing Fluoroscopy-Guided Procedures with Neural Network-Based Deformable Organ Registration

Thèse de doctorat de l'Université de Strasbourg

École doctorale n° 269, Mathématiques, Sciences de l'Information et de l'Ingénieur, MSII
Spécialité de doctorat: Informatique et Mathématique
Unité de recherche: Laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie (ICube) UMR 7357

Thèse présentée et soutenue à Strasbourg, le 18 décembre 2024, par

François LECOMTE-DENIS

Composition du jury

Su RUAN Professeure, Université de Rouen, Laboratoire LITIS, Rouen	Présidente
Marie-Odile BERGER Directrice de Recherche, INRIA, Équipe TANGRAM, Nancy	Rapportrice
Yipeng HU Professeur, University College London, Londres	Rapporteur
Su RUAN Professeure, Université de Rouen, Laboratoire LITIS, Rouen	Examinatrice
Nazim HAOUCHINE Assitant professor, Surgical Planning Laboratory, Harvard Medical School, Boston	Examinateur

Direction de la thèse

Stéphane COTIN Directeur de Recherche, INRIA, Équipe MIMESIS, Strasbourg	Directeur
Jean-Louis DILLENSEGER Professeur, Université de Rennes, Rennes	Codirecteur

Acknowledgements

First of all, I would like to sincerely thank Marie-Odile Berger, research director at Inria Nancy and Yipeng Hu, professor at University College London, for taking the time to review this thesis work. I really appreciated their comments, which have been very helpful and constructive. I am also grateful to Su Ruan, professor at University of Rouen, for her involvement in both my PhD follow-up committee and as a president of the jury, and to Nazim Haouchine, assistant professor at Harvard Medical School, for accepting to be part of the jury and for the insightful discussions we had during my PhD.

During my master's degree in physics, specialized in medical imaging, I was rather unsure about my future. I knew that theoretical physics was not for me, but I was not sure which way I wanted to go. However, I was always quite interested in computer science, and I had just heard about deep learning recently. Thus, I decided to research and experiment with simple deep learning models, improving my programming skills in the process. And then, during the last semester of my master's degree, I was recruited by Stéphane Cotin for a master internship on deep learning about 2D-3D registration on medical images.

This had a profound impact on me, because I finally had a concrete way of contributing to research, in a field that could directly improve the lives of patients. This internship transformed into a PhD position in the MIMESIS team at Inria, under the supervision of Stéphane Cotin, and the co-supervision of Jean-Louis Dillenseger at the University of Rennes. For this, I would like to thank Stéphane for his continued trust and support since my first day at MIMESIS, back in 2020. He has been an exceptional PhD advisor, always giving me the freedom to pursue my own ideas, suggesting relevant improvements to the method, and has helped me greatly in writing publications. Thanks to our collaboration, we have been able to publish 3 papers in international conferences, and attend these conferences in person, leading to unforgettable experiences abroad. With one more journal paper to be published, cementing this work in an exhaustive publication, I am very proud of the work we have done together.

This work would of course not have been possible without large amounts of teamwork with our clinician collaborators, Juan Verde and Simon Rouzé. Juan,

clinical researcher at the IHU of Strasbourg, has first been a great source of clinical knowledge, but he has also suggested numerous ways to improve the method, and experimental protocols to test it, which is why I would like to sincerely thank him for his invaluable help. I did not have the pleasure to collaborate as much with Simon, surgeon at the Rennes CHU, as I would have liked, but thanks to him, I could have a glimpse into real clinical practice, and I could even attend a real image-guided intervention on a patient, which was a very enriching experience. He also strived to provide us with data and worked on a protocol for an experimental protocol in collaboration with the IHU, even though administrative constraints did not facilitate our work together.

I would also like to thank Pablo Alvarez for our stimulating and numerous discussions, and for his help on the integration of biomechanics in the method. Since we first met, before your arrival at MIMESIS, you have always been very supportive and helpful, taking time to explain things and discuss with me, and I am very grateful for that. It has been a great pleasure to work and climb after work with you, Pablo.

Thanks to you, Valentina Scarponi, my work on registration demonstrated its potential in combination with your autonomous endovascular navigation method, and our collaborative paper allowed us to attend IROS 2024 in Abu Dhabi, where we could even make friends with some camels ! You were always kind and mindful of others, and also a perfectionist, and for that I really appreciated working with you.

I had the chance to work in a very friendly team, and for that I am grateful to all my colleagues in the lab, with whom I shared not only an office, but also lots of fun and great memories. Thank you MIMESIS and MLMS team !

A special mention goes to my friends, Pierre, Nacer, Adèle, Améline, and all the others, first for the good times we had together, but also for helping me to grow as a person and for all the support they have given me.

Finally, I want to express my gratitude to my family, who believed in me and supported me in my journey. I would like to deeply thank my parents, who have showered me with love, have always been there for me, and unconditionally accepted me as I am. Anne and Jean-Pierre, you have my deepest gratitude for accepting me in your family and your selfless love and support. My final word is dedicated to the love of my life, my husband Louis, who put up with me during my PhD, and without whom I would certainly not have had the strength to go so far. I am incredibly lucky to be at your side and for your love and tenderness.
I love you.

Table of contents

Chapter 1

Context & introduction

1.1 Benefits and challenges of image-guided interventions

In the last decades, the improvement of medical imaging systems and computers has led to the development of image-guided interventions. In image-guided interventions, the clinical team uses medical imaging systems to obtain a detailed image of the patient's internal anatomy. Before the intervention, preoperative images are acquired to plan the surgery. The images are often 3D scans of the patient, offering a very detailed view of the anatomy and enabling the localization of important anatomical structures. Due to the accuracy, often sub-millimetric, and the level of detail 3D scans offer, they are now routinely used to plan interventions (Cleary *et al.*, 2010). In parallel, intraoperative imaging systems allow clinicians to see and operate on the internal anatomy of the patient without having to create large incisions.

Both diagnostic imaging and intraoperative imaging had a positive impact on clinical practice. In abdominal surgery, intraoperative imaging has made minimally invasive surgery possible, resulting in fewer complications and reduced postoperative hospital stay, improving safety and quality of care for patients (Alkatout *et al.*, 2021; Falcoz *et al.*, 2015; Buia *et al.*, 2015). In neurosurgery, preoperative and intraoperative Magnetic Resonance Imaging (MRI) has significantly improved the visualization of brain tissue for surgeons, leading to better patient outcomes (Hall *et al.*, 2003). In orthopedic surgery, fluoroscopic and Computed Tomography (CT) imaging have improved navigation, reducing reoperation rates (Dea *et al.*, 2016).

The clinical workflow of image-guided procedures is variable, but in many cases, a 3D CT or MRI scan of the patient is acquired as the first step of the workflow to plan the intervention. From the preoperative scan, the clinical team localizes the anatomical sites to operate on and the location of incisions to be performed.

Then, during the intervention, imaging systems are used to guide the clinicians by providing information about the location of anatomical structures and surgical instruments in the operating field. Interventional imaging systems include CT, Cone-Beam CT (CBCT) and MRI for 3D imaging, and fluoroscopy, ultrasound, and endoscopy for 2D imaging.

The choice of imaging modalitie(s) depends on the clinical application. For example, in abdominal endoscopic surgery, several surgical ports (circular openings a few centimeters in diameter) are created in the patient's abdominal wall. The clinical team visualizes the anatomy using an endoscopic camera inserted in one port, while the other ports are used to insert surgical instruments and perform the operation. Due to the indirect and partial view of the anatomy and the restricted surgical openings, endoscopic surgery requires specific training.

Other interventional imaging modalities such as ultrasound, MRI, and X-Ray based imaging, do not require surgical ports to visualize the patient's anatomy. Thanks to these modalities, procedures that do not require incisions have been developed. Examples of such procedures include percutaneous procedures, where needle-like tools are inserted through the skin to operate, and angiographic procedures where a catheter is inserted in the vascular system to reach the treatment site. These modalities are also often used to localize important anatomical structures, such as tumors, before an incision is performed (Rouze *et al.*, 2016). With each interventional imaging modality presenting different characteristics, the choice of modality for a given procedure takes into account factors such as cost, procedure length, patient and clinician radiation exposure, and ease of performing the procedure.

As introduced above, endoscopic camera guidance is suited to minimally-invasive interventions because it offers a direct, optical, view of the patient internal organs as in conventional, open surgeries. Drawbacks of endoscopic camera guidance include: surface-only view of the anatomy, limited field of view, and the requirement of a dedicated camera operator besides the surgeon (Tonutti *et al.*, 2017).

Ultrasound guidance is widely used due to its efficacy and low cost (Kumar *et al.*, 2009; Harvey *et al.*, 2012; Bisset *et al.*, 2012). Nevertheless, this modality presents challenges in specific cases: ultrasound imaging of the lung is limited due to the presence of air in the organ, which blocks the transmission of the ultrasound beam (Douglas *et al.*, 2001). In the liver, heterogeneity of the parenchyma, among other factors, limits the visibility of tumors in ultrasound images (M. W. Lee *et al.*, 2010; Puijk *et al.*, 2018). Other common challenges for ultrasound guidance include poor visibility, limited field of view (Noble *et al.*, 2011), and operator-dependent image quality (Findl *et al.*, 2003; Glor *et al.*, 2005).

MRI is a 3D imaging modality, often used in neurosurgery and increasingly in cardiovascular procedures (Campbell-Washburn *et al.*, 2017). Despite its non-

ionizing nature and high image quality, its widespread use in interventional settings is still limited by its high cost, difficulty in operating inside the MRI scanner and its requirement for non-ferromagnetic surgical equipment (Manhire *et al.*, 2003).

Another 3D imaging modality, CT scan, has been employed in interventional settings, notably for biopsies. CT-guided biopsy is an incremental procedure wherein a biopsy needle is percutaneously advanced in millimeter intervals, with CT image acquisition at each step to verify needle trajectory relative to the target lesion. CT-guided biopsies are inherently time-consuming processes because clinicians have to leave the operating room during CT acquisition due to ionizing radiation (Sarti *et al.*, 2012). Furthermore, a large number of CT scans acquisitions over the course of treatment may result in an increased cancer risk for the patient, warranting a limited use of this modality in clinical practice (McCollough *et al.*, 2015).

Fluoroscopy, similarly to CT, is an ionizing imaging modality, where X-Rays are emitted from a source, traversing the patient before depositing their energy in a detector. In modern fluoroscopy imaging devices, the detector is a solid-state, position-sensitive, detector that converts the X-Ray photon energy deposited in a pixel into an electrical current, much like a digital camera. In CT scans, hundreds of X-rays 'pictures' of the anatomy are taken to produce a 3D image, while fluoroscopy provides clinicians with an instantaneous 2D image of the anatomy. In fluoroscopy, it is also possible to acquire a 'video' of the anatomy at a rate of ~ 15 Hz for real-time guidance, at the cost of elevated radiation doses for the patient and clinicians. In cardiovascular interventions, real-time fluoroscopic imaging is an indispensable tool, providing clinicians with a high-resolution view of the vascular network via contrast agent injection (Celi *et al.*, 2017). Dependence on contrast agents to visualize vessels is a drawback of X-ray based modalities due to the notable nephrotoxicity and potential of anaphylactic shock caused by contrast agents (Mantz *et al.*, 1982; McClennan, 1990; Y.-W. Wu *et al.*, 2016). Furthermore, some fine structures such as tumors or organ boundaries may not be visible in fluoroscopic images, limiting the usefulness of the modality. Finally, when operating under fluoroscopy guidance, the clinical team is more exposed to radiation since they do not leave the room during acquisition, contrary to CT guidance, with non-negligible effects on the clinician's health (Gislason-Lee *et al.*, 2016).

1.2 Existing solutions to enhance image-guided interventions

First of all, preoperative planning scans of the anatomy remain under-used in 'conventional' image-guided procedures. Typically, one imaging modality is used

before intervention for diagnostic and planning purposes, and another modality is used for intraoperative guidance, but information from preoperative and intraoperative images are not combined. While preoperative data is accessible during the procedure, pre- and intraoperative images still have to be combined mentally by the clinicians (Maybody *et al.*, 2013). This is due to the necessity of both images to be aligned, or *registered*, to be combined properly. Formally, registering two images requires finding a pixel-to-pixel transform from one image to the other such that objects in one image become aligned with objects in the other image. In general, images are acquired from different points of view, and thus rigid alignment (i.e. rotation, translation, and scaling) between images is a necessary first step. Additionally, the anatomy of the patient may be deformed from one image to another (e.g. due to breathing), requiring deformable registration to be performed. Finally, the modality of both images may be different, further complicating the registration task.

Secondly, relating the content in interventional images to the surgical scene is not straightforward either because interventional images are typically viewed on a screen above the operating table. This setup requires clinician to perform mental registration again between the operating field and the images displayed on the screen. Thanks to developments in the field of computer-aided interventions, solutions have been developed to alleviate these difficulties, some of them currently being used in operating rooms.

The SimpliCT (NeoRad BV, Netherlands) laser guidance system, developed to facilitate needle insertion, represents one such computer-aided navigation technology. In (Varro *et al.*, 2004), Varro *et al.* detail the interventional use of SimpliCT. The SimpliCT device is a tripod-mounted laser source that is first aligned with the CT laser guidance. Then, a CT scan is acquired to determine the optimal needle insertion angle and the SimpliCT guidance laser is set to the determined angle. Finally, the needle is inserted following the angle given by the SimpliCT device under suspended breathing. Cited benefits of this product are shorter procedure times and reduced irradiation (Varro *et al.*, 2004; Kroes *et al.*, 2016). Since no registration is performed by this product, its use is limited to static anatomies, for example requiring suspended breathing in abdominal interventions (Varro *et al.*, 2004).

1.2.1 Augmented and Virtual Reality

With the development of wearable displays for virtual and augmented reality, applications of these technologies in the operating room have been investigated, and several reviews have been published (Yoon *et al.*, 2018; Z. Zhao *et al.*, 2021). These displays are in the form of glasses or screens worn in front of the eyes, such that the user is always looking at the display. Augmented reality (AR) refers to

transparent displays that allow the user to view the real world, augmented with superimposed digital images, while virtual reality (VR) displays show either a fully virtual scene or a live camera feed of the real world augmented with digital images. A straightforward application of these technologies in the operating room is to augment the clinician's view with information that they would otherwise view on a separate display.

In anesthesia, it has been reported that using a wearable display to monitor the patient vital signs shortened the time to detect risky situations in simulated (Ormerod *et al.*, 2003; Przkora *et al.*, 2015) and real (D. Liu *et al.*, 2010) interventions. The MicroOptical (MicroOptical Corporation, USA) head-mounted display has been developed for orthopedic surgery, to view fluoroscopic images in augmented reality rather than on the fluoroscopy monitor. In a study (Ortega *et al.*, 2008), Ortega *et al.* found that, when using the MicroOptical display, the surgeon left the attention of the operative field to view fluoroscopic images less often (5 times) than without (207 times). The same study, however, found no noticeable reduction in fluoroscopy time and noted the tripping hazard of the cable connecting the MicroOptical display to the fluoroscopy monitor. The Google Glass (Google Inc., USA) product has been used in several studies covering different clinical applications (Yoon *et al.*, 2018). In most application, this head-mounted display was used to display various information in augmented reality. In cardiothoracic transplantation surgery, Google Glass was used to livestream recovery of a lung for transplant on a patient in a separate location. The transplant clinical team could evaluate in real-time the organ quality and anatomical suitability, in coordination with the recovery clinical team. The study authors noted the potential benefits of the Google Glass' live streaming functionality in transplant surgery (Baldwin *et al.*, 2016).

In percutaneous interventions, clinicians have no direct view of the anatomy and must rely on fluoroscopy, CT, US, or MR images to ensure the needle is inserted at the correct position and angle. Traditionally, this image guidance is viewed on a separate screen and is not augmented with preoperative information. Thus, AR has the potential to improve percutaneous interventions by providing the surgeon with anatomical information superimposed on the operative scene.

In two studies (De Paolis and Ricciardi, 2018; De Paolis and De Luca, 2019), an AR headset was used to visualize a preoperative 3D segmentation of the liver, complete with the vessels and the tumor, superimposed on the patient anatomy for percutaneous Radiofrequency Ablation (RFA) of liver tumors. The 3D preoperative data was rigidly registered to the intraoperative anatomy thanks to radiopaque fiducial stickers visible in the preoperative CT scan and on the patient skin. An optical tracker system, consisting of four infrared cameras, is used to localize the position of the fiducials in the intraoperative scene for rigid registration of the

CT data on the intraoperative scene. The authors noted that deformable motion of the anatomy may occur in abdominal surgery, due to breathing and needle-tissue interactions, which would invalidate the augmented visualization, but did not implement a solution for this issue. Other studies have also investigated the use of AR for percutaneous interventions, with a similar lack of deformable registration to take into account organ motion (Solbiati *et al.*, 2018), (Park *et al.*, 2020)[1]. Finally, in (Kuzhagaliyev *et al.*, 2018), Kuzhagaliyev *et al.*[1] evaluated the use of AR, with the Hololens (Microsoft, USA) platform, to plan and guide needle insertion for tumor ablation in the pancreas. In this procedure, an ultrasound probe is used to localize the tumor and define the needle trajectory. After registering the US probe, needle, and Hololens in a common coordinate system using infrared cameras, the authors showed, as a proof of concept, that real-time ultrasound images of the anatomy could be overlaid on the patient for AR ultrasound-guided interventions. Notably, in this case, only rigid registration is necessary because the ultrasound probe produces images of the anatomy in real-time during the intervention.

1.2.2 Solutions for bronchoscopic interventions

In contrast with wearable displays, still scarcely used in clinical practice, screens are ubiquitous in image-guided interventions, being used to monitor vital constants and display preoperative and intraoperative data. Thus, solutions have been developed and commercialized to augment information displayed on screens, with positive impacts on procedure time, especially for less experienced clinicians (Detmer *et al.*, 2017; Mert *et al.*, 2012; Vles *et al.*, 2020).

In the lung, Endobronchial Ultrasound (EBUS) is a widely used imaging modality to identify lung nodules to be diagnosed, by navigating an ultrasound probe through the bronchia. After identifying nodules, a biopsy is performed by navigating a bronchoscope, a type of endoscopic camera, in the bronchia up to the nodule site. Since the EBUS images are not registered to the bronchoscopic images, navigating to the nodule is not trivial.

To solve this problem, Olympus (Tokyo, Japan), commercialized a virtual bronchoscopic navigation system that uses the preoperative CT scan to build virtually navigable bronchia. During the intervention, one clinician navigates the real bronchoscope in the patient's anatomy while another clinician simultaneously navigates a virtual bronchoscope in the preoperative anatomy. Thanks to the synchronous navigation in the real and virtual environments, clinicians can approximately localize the position of the bronchoscope with respect to the preoperative CT anatomy. One study by Ishida *et al.* (Ishida *et al.*, 2011) reported a higher diagnostic sensitivity with virtual bronchoscopic navigation than without. In another, earlier, study,

[1] Arxiv pre-print

(Fumihiro Asano *et al.*, 2006), Fumihiro Asano *et al.* also reported a high diagnostic sensitivity for virtual bronchoscopic navigation. Limitations of this method are the requirement of a dedicated operator for the virtual navigation system and the possibility of desynchronization between the virtual and real scenes, leading to erroneous navigation (Asano *et al.*, 2002).

The Archimedes (Broncus Medical, China) system is another, more recent, product developed for bronchoscopic navigation. Although no work has precisely detailed how bronchoscopic navigation is performed using Archimedes, the general navigation process has been presented in several studies (Q. Zhang *et al.*, 2021; Sun *et al.*, 2022; Lanfranchi *et al.*, 2024). At the start of the intervention, Archimedes automatically computes the path in the bronchia to the lesion from a preoperative CT image. During the intervention, the preoperative data is rigidly registered to the intraoperative anatomy with the help of a positioning board. To reduce the mismatch between preoperative and intraoperative anatomies, the patients are put under volume-controlled ventilation. Then, a bronchoscope is matched with the Archimedes system and navigated to the lesion following preoperative planning with the help of augmented interventional imaging. In an early study, (Q. Zhang *et al.*, 2021), Q. Zhang *et al.* evaluated the efficacy of Archimedes for transbronchial cryobiopsy. Although the sample size of the study is small (8 patients), the authors confirmed that Archimedes could be used for navigation in transbronchial cryobiopsy. Another study by Sun *et al.* (Sun *et al.*, 2022) and sponsored by Broncus Medical, investigated the use of Archimedes for bronchoscopic transparenchymal nodule access (BTPNA) and transbronchial needle aspiration (TBNA) on 104 patients. In these procedures, radial EBUS and fluoroscopic images augmented with a visualization of the tumor were used for guidance. Specific details on the use of Archimedes and the augmentation of fluoroscopic images were missing, but the authors reported a high diagnostic sensitivity and acceptable procedure durations. Finally, (Lanfranchi *et al.*, 2024), Lanfranchi *et al.* also evaluated Archimedes for BTPNA and TBNA, and reported results consistent with previous studies.

The LungVision (Body Vision Medical, Israel) system is a CT and fluoroscopy-based navigation solution developed for bronchoscopic interventions. The system uses dynamic fluoroscopic acquisition during C-arm rotation to reconstruct an intraoperative 3D scan of the patient, a process denoted C-arm Based Computed Tomosynthesis (CABT). This is an advantage over other solutions relying on more expensive CBCT devices, as C-arms are more readily available in interventional suites. Notably, the system makes use of a positioning board embedded with radio-opaque markers for tomosynthesis and tracking purposes. In (Bawaadam *et al.*, 2024), Bawaadam *et al.* detail the clinical workflow of bronchoscopic navigation with LungVision for peripheral lesion biopsy in the lung. First, a series of fluoroscopic images is acquired to register the intraoperative scene to the pre-

operative CT. Bronchoscopy equipment is registered to the preoperative anatomy during this step as well. In this way, the lesion is approximately localized in the intraoperative scene and the C-arm can be re-positioned to center it on the lesion. Then, a CABT scan is acquired to precisely localize the lesion in the intraoperative anatomy. Since the positioning board is visible in both the CABT and fluoroscopic images, it is possible to correct for motion of the C-arm with respect to the table. Navigation inside the bronchia following the path defined in the preoperative CT is then performed. Once navigation is completed, another CABT scan is acquired, with the bronchoscopy tools in view, to precisely localize the tools with respect to the lesion. Then, tools are advanced to the lesion and, optionally, a final CABT image is acquired to confirm the position of the tools in the lesion. Finally, the biopsy is performed under fluoroscopic guidance. (Bawaadam *et al.*, 2024), Bawaadam *et al.* cite several studies (Pritchett, 2021; Cicenia *et al.*, 2021; Aboudara *et al.*, 2020; Wagh *et al.*, 2021; Hedstrom *et al.*, 2022; Pertzov *et al.*, 2021) on the lesion localization and diagnostic yield performances of LungVision, which are comparable with the previously cited Archimedes system.

The Olympus, Archimedes, and LungVision systems cited above have demonstrated clinical improvements in bronchoscopic interventions. Olympus, which proposes an innovative virtual navigation system, effectively leverages the preoperative CT scan to facilitate navigation, although the application of this concept to other types of interventions remains unclear. In contrast, the Archimedes system's solution to the problem of preoperative to intraoperative rigid registration through the use of a positioning board could readily be applied to other types of interventions, even though specific details on the approach are missing. Finally, the LungVision system demonstrates the benefits of CBCT-fluoroscopy, which is an effective and broadly applicable interventional guidance solution, as detailed below.

1.2.3 CBCT-fluoroscopy navigation

The development of CBCT imaging, combining fluoroscopy and CT capabilities in a compact device, has made intraoperative augmented fluoroscopy possible. CBCT imaging devices track the position of the detector and X-ray source relative to the operating table at all times. Thus, the system can reconstruct 3D CBCT images from a series of fluoroscopic images acquired during rotation, similarly to conventional CT acquisition. Furthermore, this internal coordinate system also allows fluoroscopic images to be augmented with a previously acquired 3D CBCT scan, since the pose at which the fluoroscopic image was acquired is recorded in the system. Numerous clinical workflows have employed this capability: in (Schafer *et al.*, 2020, pp. 654-663), Schafer *et al.* report that combined CBCT-fluoroscopy imaging is used for interventions in the vessels of the brain, abdominal organs, and

the heart, and in orthopedic interventions, tumor ablation procedures, as well as radiotherapy applications.

For example, in angiographic interventions, where the clinical team inserts a catheter to operate inside the vessels, fluoroscopy is used to guide the catheter insertion and visualize the vessels via contrast agents injections. However, contrast agent injections are inherently limited due to blood flow dissipating contrast quickly and nephrotoxicity. Consequently, in practice, parts of the navigation are performed without contrast, with the help of a roadmap of the vessels. Because registration between intraoperative 2D and preoperative 3D images is still an experimental technology, it is not common clinical practice to use a preoperative roadmap. Thus, the roadmap is obtained intraoperatively via contrast agent injection, either using 2D fluoroscopic acquisition or 3D CBCT acquisition. In this case, CBCT imaging provides clinicians with a 3D reconstruction of the vessels, enhancing the visualization of the anatomy and helping navigation. Afterward, fluoroscopic images can be augmented with the 3D roadmap rigidly superimposed on the 2D images to help navigation. In addition to improving the interventional workflow, intraoperative CBCT images may bypass the need for preoperative CT scans in time-critical interventions such as ischemic stroke treatment (Maier *et al.*, 2018).

Percutaneous interventions represent another type of intervention where CBCT imaging is commonly used for guidance. In this type of intervention, the clinician introduces a needle-like tool under image guidance to reach a target structure. Here, the clinician's objective is twofold: accurately reach the target structure and minimize damage to healthy tissues. Although the imaging modality utilized in percutaneous interventions varies, CBCT imaging has been employed in interventions such as Transjugular Intrahepatic Portosystemic Shunt placement (TIPS) (Ketelsen *et al.*, 2016), ultrasound-guided microwave ablation (Floridi *et al.*, 2017), transthoracic needle biopsy (Choi *et al.*, 2012), and other percutaneous tumor ablation procedures in the liver, kidney, lung, and muscles (Abi-Jaoudeh *et al.*, 2015). In these interventions, a CBCT image may be acquired at several points during the intervention to accurately assess the position of the tool relative to the target and surrounding structures. Additionally, the CBCT image can also be used to define a needle insertion trajectory at the start of the intervention. Then, the clinician can rely on the augmented fluoroscopy guidance showing the insertion trajectory and relevant anatomical structures to insert the needle with improved effectiveness.

While intraoperative CBCT guidance represents an improvement over fluoroscopy-only guidance, a single CBCT acquisition is equivalent in dose to several minutes of continuous fluoroscopic acquisition (Sailer *et al.*, 2015). Thus, if a preoperative roadmap is available, it would be beneficial to register it to intraoperative fluo-

roscopic images to reduce radiation doses. Furthermore, at present, the accuracy of augmented fluoroscopic guidance is constrained by deformations that arise during needle insertion, as there are no comprehensive methods for general 2D/3D deformable registration. These deformations lead to discrepancies between the overlaid CBCT data and the fluoroscopic image, as noted in (Wallace *et al.*, 2008).

1.3 Our approach for enhanced fluoroscopy-guided interventions

We focus here on fluoroscopy-guided interventions and, more specifically, seek to enhance the information content of fluoroscopic images.

Fluoroscopic images are high-resolution images showing the full internal anatomy in their field of view, that suffer from low or absent contrast for important anatomical structures such as vessels or tumors. In contrast, these structures are often visible and segmented in preoperative 3D CT scan images. Consequently, information from preoperative images could be used to enrich the intraoperative images, for example by superimposing the preoperative data on intraoperative images. However, superimposing preoperative data on interventional images requires a non-linear transformation from the 3D preoperative image space to the 2D intraoperative image space. Because the correspondence between interventional imaging systems and preoperative image systems is unknown, we must first find the transform between the preoperative and interventional reference frames, an operation defined as rigid registration. In the specific context of laparoscopic surgery, breathing and surgical motions induce a deformation of the abdominal organs. Recovering this deformation is necessary to obtain the correct transformation between preoperative data and the intraoperative anatomy, an operation known as deformable registration. Thus, to propose a solution for augmented fluoroscopy-guided interventions, we developed a fluoroscopy-to-CT deformable registration method. Our method offers several key technical and clinical advantages:

- Operates with standard C-arm fluoroscopy equipment, eliminating the need for additional specialized imaging hardware or intraoperative CBCT acquisition

- Requires only a single preoperative 3D CT scan, avoiding the complexity and increased radiation exposure of a 4D CT acquisitions

- Has the potential to enhance patient safety by eliminating the need for intraoperative contrast agent injection, reducing procedure-related risks while maintaining visualization quality

- Represents the first fluoroscopy-to-CT registration method designed for interventions beyond radiotherapy, significantly expanding its clinical applications

1.3.1 Thesis outline

Chapter 2 begins with an overview of the different types of image registration approaches, outlining their associated challenges. Then, we introduce deep learning principles as applied to image registration, presenting a concise review of foundational works in this domain.

Chapter 3 establishes the current state of the art through a comprehensive review of methods related to our work. This chapter is organized into four key sections. First, section 3.1 examines tumor localization approaches, aiming to detect rather than register structures in fluoroscopic images. Section 3.2 covers 2D-3D rigid registration methods, which form an essential preliminary step in any registration pipeline. Section 3.3 examines fluoroscopy-to-CT deformable registration methods, which represent the approaches most closely aligned with our work. These methods, like ours, typically leverage an initial rigid registration as their foundation. Finally, section 3.4 examines registration methods that integrate biomechanical models as a strategy to both handle incomplete information and constrain solutions to physically realistic deformations.

Chapter 4 introduces our framework (section 4.2) and presents a series of experimental validations through published and unpublished works. In section 4.3, a preliminary study validates our data generation approach, departing from traditional statistical deformation models-based methods. Section 4.4 reproduces our first published work, presented at the Hamlyn Symposium on Medical Robotics (HSMR), demonstrating how our method can replace contrast agent injection for vessel visualization in fluoroscopy-guided interventions. This is followed by section 4.5, which details a comprehensive study submitted for publication in the Medical Image Analysis journal, validating our method's capability to recover intervention-related deformations. The chapter concludes with section 4.6, describing our work presented at the International Conference on Intelligent Robots and Systems (IROS), where our method is combined with the method of Valentina Scarponi, a fellow team member, for fluoroscopy-guided autonomous catheter navigation.

Chapter 5 extends our methodology to incorporate physics-based constraints, enhancing registration realism. The first section presents a study on physics-based data generation for realistic deformation prediction, presented at the Data Curation and Augmentation in Enhancing Medical Imaging Applications (DCAMI) Workshop of the Computer Vision and Pattern Recognition (CVPR) conference. In a second section, we present experiments on the incorporation of physical regu-

larization during training to further improve predictions' realism. In these experiments, we study different architectural variants of our network and present results on a simplified, 2D registration problem.

Chapter 6 validates our method on real fluoroscopic images through two experiments. The first evaluates 2D accuracy during respiration using implanted landmarks in a porcine model, while the second assesses our method's clinical applicability for tumor localization on clinical data.

Finally, Chapter 7 concludes the manuscript by summarizing our contributions and discussing future directions for improving and validating the clinical utility of our method.

Chapter 2

Image registration

Image registration is the process of establishing a spatial correspondence between a fixed and a moving image. The aim of the registration algorithm is to find the best spatial correspondence according to some criterion (e.g. the overlap of corresponding anatomical structures). This matching is typically modeled by a function of coordinates between the fixed and moving image domains. Warping the moving image with the registration transform makes it possible to superimpose elements of the moving image onto the fixed image.

In the medical field, it is common to acquire several images of the same patient, at different points in time. For example, several diagnostic scans of the anatomy may be acquired over the course of months to years in order to study the evolution of a pathology. Because patient positioning vary between imaging sessions, rigid registration must first be performed to establish a global alignment between different image acquisitions. This initial alignment also compensates for changes in position, orientation, and scale between the imaging coordinate systems, providing a necessary foundation for subsequent registration steps. In addition, breathing, bowel movements, cardiac motion, or other sources deform the patient's internal anatomy, requiring deformable registration to fully align images. Once the images are well registered, their contents can be fused to enable a more complete visualization of the anatomy.

2.1 Image registration: an overview

2.1.1 Unimodal and multi-modal registration

Image registration methods can be split into two broad categories: unimodal registration, where the imaging modality is the same for all images, and multimodal registration, where two or more imaging modalities are used across images. Uni-

modal registration is the most common type of image registration, with works dating as far back as 1977 (Van den Elsen *et al.*, 1993; Barrow *et al.*, 1977). In unimodal registration, intensity values represent the same information in both images, and the registration accuracy can be evaluated by directly comparing the registered image with the target image. This type of registration is often used to study the evolution of the anatomy in time.

In multimodal registration, the content, dimensionality and image formation process can vary across images. Thanks to the differences in image formation processes, registering images of different modalities brings complementary information to clinicians. For example, in MRI, T2*-weighted images are used to visualize fluids and bones, and STIR images are used to visualize tumors or inflammation (Brown *et al.*, 2011). Thus, registering a T2*-weighted and a STIR image allows to visualize fluids, bones, tumors and inflammation on the same image. An example of the difference in information content between modalities is illustrated in Fig. 2.1, where all images were rigidly aligned to combine Cerebral Blood Volume (CBV) and Apparent Diffusion Coefficient (ADC) with T1 and T2 images (Wuerfel *et al.*, 2004).

Fluoroscopy to CT registration is an example of multimodal registration between images of different dimensionalities. In this case, the preoperative, information-rich 3D image complements the real-time, intraoperative 2D image. However, while it is desirable to enhance intraoperative 2D images with rich preoperative 3D information, it is also more difficult, owing to the combined modality and dimensionality difference.

2.1.2 Rigid and deformable registration

As previously mentioned, the image registration process is split into two parts. The first part, rigid registration, consists of finding a global coordinate transform that aligns the origin, orientation, and scale of both images. This is best understood by looking at Fig. 2.2, where the origin, orientation, and scale alignment operations are illustrated. Mathematically, rigid registration is modeled as an affine transform between the moving image space and the fixed image space. It is possible to represent this transform in matrix form as such:

$$\begin{bmatrix} \mathbf{y} \\ 1 \end{bmatrix} = \left[\begin{array}{c|c} R & \mathbf{T} \\ 0 & 1 \end{array}\right] \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} \tag{2.1}$$

Where $\mathbf{x}$ is a point in the moving image space, R is a rotation matrix, $\mathbf{T}$ is a translation vector and $\mathbf{y}$ is a point in the fixed image space. Notice that $\mathbf{x}$ and $\mathbf{y}$ are expressed as homogeneous coordinate vectors, with an additional component equal to 1 at the end.

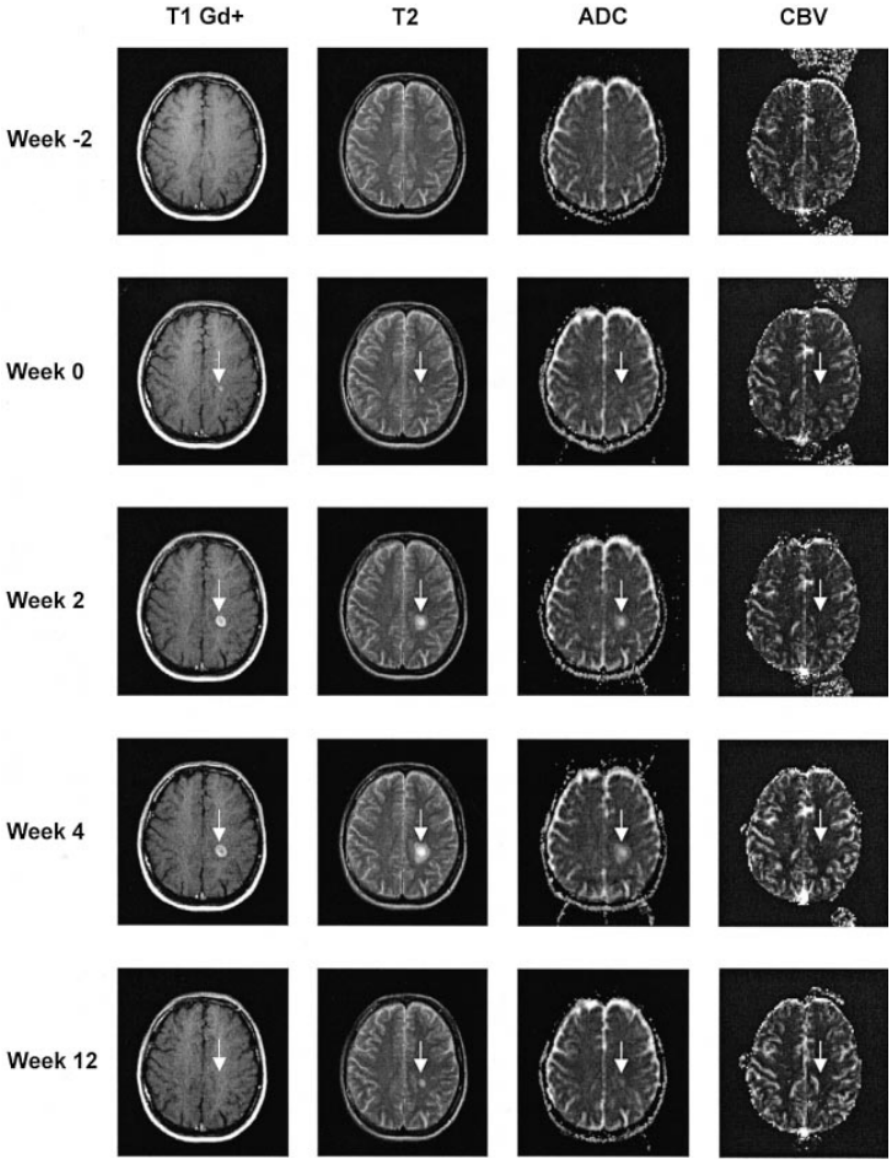


Figure 2.1: From left to right, T1-weighted, T2-weighted, ADC and CBV MRIs of a patient brain. Each image is acquired with a different process and shows different information. This is evidenced by the lesion under the arrow, which is more or less visible depending on the modality, even though all images in a row were acquired at the same point in time. This figure was reproduced from (Wuerfel *et al.*, 2004), with publisher's permission.

In general, rigid registration is necessary because different imaging devices may have different reference frames (requiring change of frame correspondence), and the

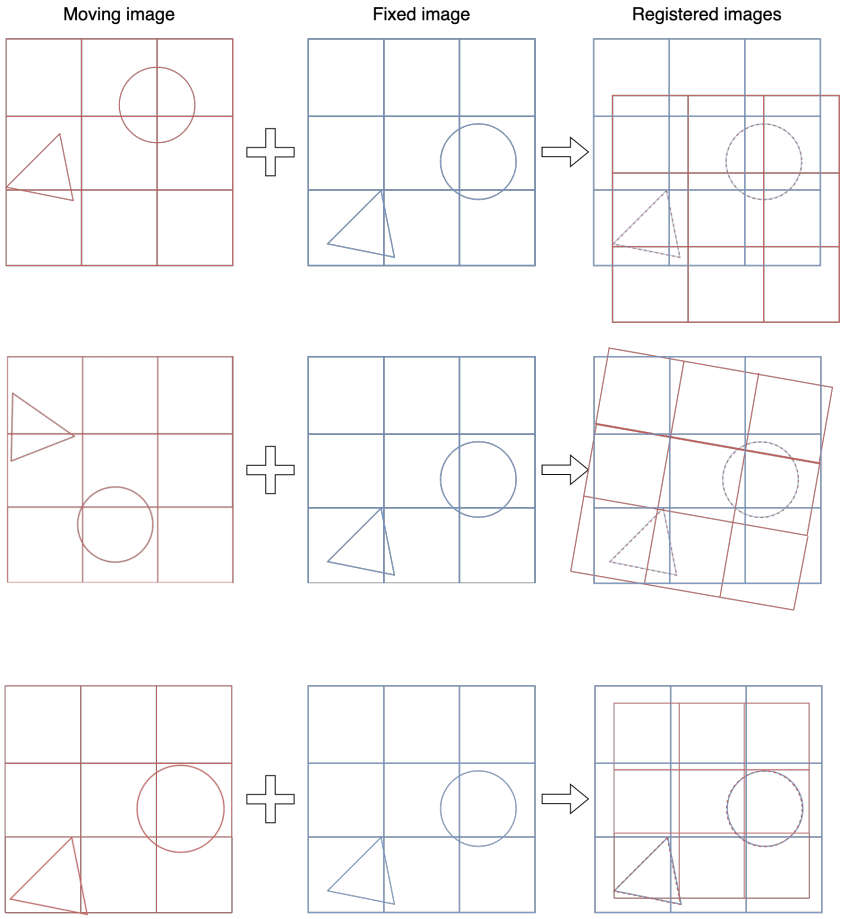


Figure 2.2: From top to bottom, origin, orientation, and scale alignment of the moving image (red) on the fixed image (blue). Note that, in general, the moving and fixed image domains are different and the contents of both images can overlap only at the intersection of both domains.

position of the patient may vary across images. Additionally, the resolution of the imaging device may change, such that the size of objects in terms of pixels is different across images.

In some specific cases, the structures of interest in the anatomy may not be deformable, like rigid bones such as the femur or the hip, and rigid registration is sufficient. However, in general, tissues are deformable to some degree and deformable registration is necessary to finely align anatomical structures between images. A diagram representing the effect of deformable registration is presented in Fig. 2.3. Contrary to rigid registration, deformable registration cannot be modeled by an affine transform. Instead, the deformable registration transform is expressed either parametrically or in a discrete form. A parametric registration

transform can be expressed as:

$$y = f(x, \lambda_1, \ldots, \lambda_n)$$

where $\{\lambda_1, \ldots, \lambda_n\}$ are transformation parameters and f a parametric function.

In contrast, a discrete registration transform is represented as an element of a high-dimensional space, similarly to an image. If the discrete transform is applied to an N dimensional image, it is represented as an $N \times D_1 \times \cdots \times D_N$-dimensional tensor, where $(D_1, \ldots, D_N)$ are the sizes of the image in each dimension.

While discrete transforms can represent any deformation, even unrealistic ones, parametric representations can prevent unrealistic deformations by construction, but are limited in accuracy if too few parameters are employed.

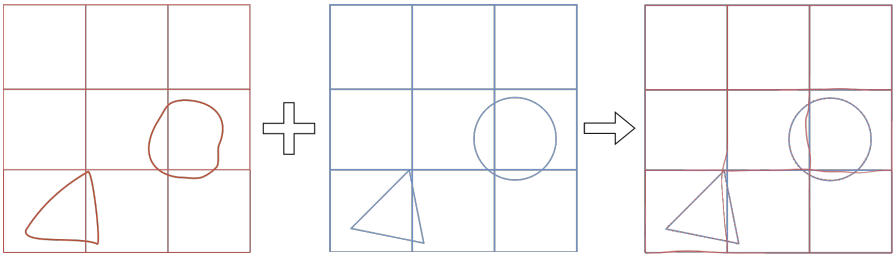


Figure 2.3: Deformable alignment of the moving image (red) on the fixed image (blue). Objects in the moving image (circle and triangle) have undergone deformations. To superimpose objects in the moving image on the fixed image, the moving image is deformed (notice the warped grid lines).

2.1.3 Biomechanical model-based registration

Biomechanical model-based registration methods are a special kind of parametric registration methods that leverage prior knowledge about the physical properties of objects to guide registration. Directly based on the laws of mechanics, these methods constrain deformations to respect physical principles, and provide physically-plausible interpolation in regions where deformation cannot be directly determined from image data.

A biomechanical model can be represented as a Partial Differential Equation (PDE) that relates the force applied on an object to its deformation, parameterized by the physical properties of the object. These properties, such as the Young's modulus E (ratio between applied force and deformation) and Poisson's ratio ν (expansion perpendicular to applied force), can be measured experimentally.

One of the simplest biomechanical models is the linear model of elasticity, represented by the Navier-Lamé equations (Modersitzki, 2003, p. 83):

$$\mathbf{F} = \mu\nabla^2\boldsymbol{\varphi} + (\lambda + \mu)\nabla(\nabla \cdot \boldsymbol{\varphi}) \tag{2.2}$$

where $\mathbf{F}$ is the force, $\boldsymbol{\varphi}$ the displacement, and (λ, μ) the experimentally measured Lamé parameters related to Young's modulus and Poisson's ratio.

This equation, derived from Hooke's law, assumes a linear relationship between $\mathbf{F}$ and $\boldsymbol{\varphi}$. However, this approximation only holds for small deformations, as experimental evidence shows the relationship becomes non-linear with larger deformations. To address this limitation, hyperelastic models, such as the Mooney-Rivlin (Mooney, 1940; Rivlin, 1948) model, were developed, incorporating additional experimental parameters to better model large deformations.

To solve these PDEs, numerical methods such as the Finite Differences (FD) method or the Finite Elements Method (FEM) are often used. These methods rely on a discrete representation of the physical world, and approximate PDEs locally using simple functions such as polynomials. While these methods can achieve arbitrary precision with sufficiently fine discretizations, this precision may require substantial computational costs.

An example of a biomechanical model-based registration procedure is to use partial information about $\boldsymbol{\varphi}$ to compute $\mathbf{F}$, where an optimization procedure (detailed in Sec. 2.1.5) is employed to tune $\mathbf{F}$ to minimize the model's energy. The physical foundation of these methods ensures that the computed deformation field follows realistic biomechanical behavior everywhere, especially in regions with poor contrast or missing data where physical principles guide the interpolation of the deformation field.

However, these methods require prior knowledge about object shape and physical properties, which may not always be available. For real-time applications, computational efficiency often necessitates using simpler models (e.g., corotational models (Felippa, 2000)) or coarser discretization, trading accuracy for speed. Despite these limitations, biomechanical model-based methods have proven successful in registration tasks, validating their ability to accurately represent anatomical deformations (Sotiras *et al.*, 2010).

2.1.4 Intensity-based and feature-based registration

Registration methods can be again divided into two main categories: intensity-based and feature-based methods. Whether a parametric or discrete representation is used, image registration involves the numerical computation of a set of parameters from the moving and the fixed images (in discrete representation, these parameters are simply the elements of the high dimensional representation).

Intensity-based methods compute the parameters of the registration transform T from the intensity values of I_F and I_M, the fixed and moving images, respectively. In intensity-based methods, a cost function L is used to compare the intensity values of the transformed image $I_M \circ T$ and I_F at a set of spatial locations. The value of the cost function is then used as a proxy to measure the registration error

and optimize the parameters of T accordingly. One advantage of intensity-based methods is that the cost function only depends on image intensity values, thus these methods do not require the extraction of specific features from the images to work.

Feature-based methods, on the other hand, rely on the prior extraction of features from the moving and fixed images. Features may be anatomical landmarks, represented as a set of points, or anatomical structures, represented as meshes. The cost function in feature-based methods is then a measure of the distance between the sets of points or meshes in the transformed image and the fixed image. In landmarks-based registration, there is a point-to-point correspondence between the landmarks in the moving and fixed image, and the cost function usually measures the point-wise distance between both sets of points. In mesh-based registration, there is usually no point-to-point correspondence between both meshes. Rather, the cost function is a global measure of the distance between meshes. Compared to intensity-based methods, feature-based methods are more independent to differences in contrast or even modality between images, because the image intensity values are only used to extract features from the images.

2.1.5 Optimization-based and learning-based registration

To compute the set of parameters $\lambda = \{\lambda_i\}$ of a transform T given the value l of a cost function L, an optimization algorithm must be used. The general form of the optimization algorithm is described by Alg. 1.

Algorithm 1 A generic optimization loop

$i \leftarrow 0$
$\lambda \leftarrow \lambda_{init}$
while $i < N$ & $l >$ threshold **do**
  $l \leftarrow L(T(\lambda), I_M, I_F)$
  $\lambda \leftarrow \text{optimizer}(l, \lambda)$
  $i \leftarrow i + 1$
end while

A commonly used optimization algorithm is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Bonnans *et al.*, 2003a), which uses the value of L and its gradient with respect to λ, $\nabla L(\lambda)$, to update λ and minimize L.

Traditionally, optimization has been used to directly compute the parameters λ of a transform T that register a moving image to a fixed image. Depending on the size of the images, the number of parameters λ_i, and the computational complexity of L, the optimization process can take from less than one second to several hours. Fast methods, like feature-based rigid registration methods, can be

near-instantaneous due to the low number of parameters involved (6 parameters for a 3D affine transform) and the low computational complexity of L. On the contrary, intensity-based, deformable registration methods can take a long time to compute, especially in 3D, due to the high computational complexity of L, which requires computing the value of $T(\mathbf{x})$ at a set points $\mathbf{x}$ in I_M, computing $I_M \circ T(\mathbf{x})$ and then finally computing the value of L.

To remediate to the high computational cost of optimizing λ for each new pair of I_M and I_F, learning-based methods have been developed for image registration. Learning-based methods aim to build a function f_θ that computes λ given a pair of images (I_M, I_F), which is called a sample. The parameters θ are computed using a dataset of images, by optimizing L over θ on the dataset, a process described in Alg. 2.

Algorithm 2 Learning-based optimization

$i \leftarrow 0$
$\theta \leftarrow \theta_{init}$
while $i < N \quad \& \quad l >$ threshold **do**
  $(I_M, I_F) \leftarrow \text{sample}_i$
  $\lambda \leftarrow f_\theta(I_M, I_F)$
  $l \leftarrow L(T(\lambda), I_M, I_F)$
  $\theta \leftarrow optimizer(l, \theta)$
  $i \leftarrow i + 1$
end while

Before the generalized use of deep neural networks, statistical deformation models (SDM) have been employed for learning-based registration (Sotiras *et al.*, 2010). In SDMs, an optimization algorithm (Alg. 1), computes a registration transform T_i between each moving image i in the dataset and a single fixed image. Then, the SDM $g_\omega(T_i)$ is defined as the function that, given a set of weights ω, returns the transformation $T = \sum_k \omega_k T_k$. In order to reduce the number of parameters ω, Principal Component Analysis (PCA) is usually performed to extract the principal components of T_i, although other methods have also been used (Sotiras *et al.*, 2010; Zhuang *et al.*, 2017). In short, the PCA algorithm finds the principal components $\tilde{T}$, each a linear combination of all T_i, that explain the most variance in the dataset. This means that, by taking a linear combination of the first k components $\tilde{T}$, T_i can be approximated with a set precision. In practice, k is chosen such that T_i is approximated with at least 90%. In some cases where the variability of T_i is low, k can be as small as 2 or 3. The SDM then becomes $g_\omega(\tilde{T}_i)$, with parameters $\{\omega_{0 \cdots k}\}$. Usually, g is simply defined as a linear

combination of $\tilde{T}$:

$$g_\omega(\tilde{T}_i) = \sum_{i=0}^{k} \omega_i \tilde{T}_i \tag{2.3}$$

After constructing the SDM, a new image can be registered to the reference image by optimizing ω_i with Alg. 1. SDM naturally simplifies the registration problem, since only a few parameters ω_i need to be optimized, and a strong prior on the nature of T is provided.

In fluoroscopy to CT deformable registration, SDMs have been constructed on experimentally acquired datasets with breathing motion for breathing motion prediction (see Sec. 3.3.1). However, an important limitation of SDMs is that T is approximated well only if it is close to the T_i in the dataset used to construct the SDM. Thus, SDMs cannot be constructed to recover deformations for which there is little to no data, like intervention-related deformations.

2.2 Deep learning for image registration

To remediate to the slowness of optimization-based methods and the weak generalization capability of SDMs, deep learning, has recently been massively used. Deep learning is a kind of machine learning method where the parameters of a deep neural network are optimized on a training dataset. A deep neural network is a composition of parametric and non-linear functions. It is usually composed of at least three layers, each formed by the composition of a parametric function and a nonlinear activation function (and optionally additional functions). The term 'neural' refers to the analogy between the network layers and neurons in the brain, with the parameters of the network representing the connections between the neurons.

2.2.1 Deep neural networks

Deep neural networks are, under some general conditions, universal function approximators (Hornik, 1991). In general, a N layer deep neural network $g_{\hat{\theta}}$ that approximates a function f is expressed as:

$$\begin{aligned} g_{\hat{\theta}} &= \sigma_N \circ g_{\hat{\theta}_N} \circ \cdots \circ \sigma_1 \circ g_{\hat{\theta}_1} \\ \hat{\theta} &= \arg\min_\theta L(g_\theta, f, X) \end{aligned} \tag{2.4}$$

With $\hat{\theta}$ the optimal parameters of g_θ under L on a dataset X. The layer i in a neural network is formed by the composition of g_{θ_i} and σ_i, along with other, optional, operations such as normalization, dropout, skip connections, ...

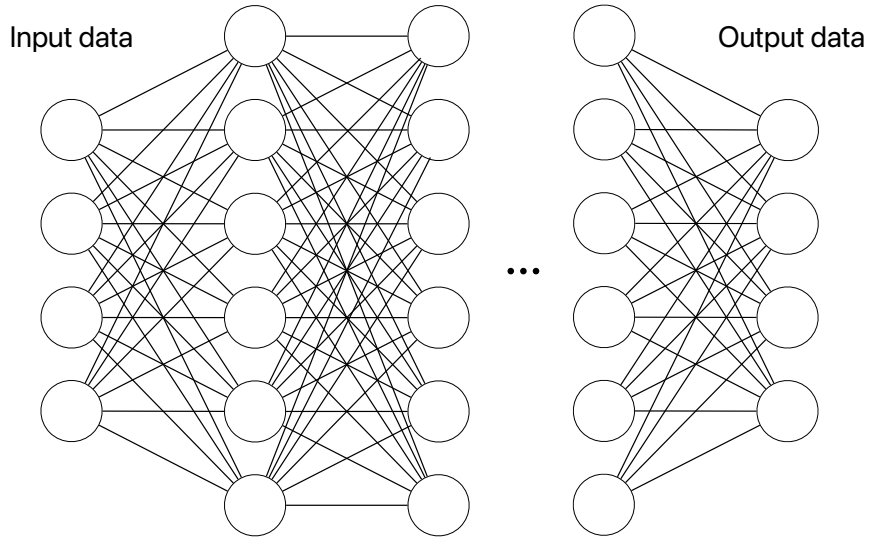


Figure 2.4: Schematic representation of a fully connected deep neural network. The circles represent the data at each layer of the network. The operation of each layer of the network is represented by the lines. With fully connected layers, each element in the output depends on each element in the input.

The functions g_{θ_i} are operators, usually linear, each acting on the previous layer output, up to the input. The functions σ_i are activation functions, non-linear by definition, acting on the output of g_{θ_i}. Commonly used operators are affine transformations $g_{\theta_i}(x) = W_i \cdot x + B_i$ and convolutions $g_{\theta_i}(x) = W_i * x + B_i$ (with $*$ the convolution operator).

There is a great diversity of activation functions, but examples of commonly used activations σ are $\mathrm{ReLU}(x) = x$ if $x > 0$ else 0, $\mathrm{PReLU}(x) = x$ if $x > 0$ else ax, and $\mathrm{GELU}(x) \approx x\frac{1}{2}(1 + \mathrm{erf}(\frac{x}{\sqrt{2}}))$. It is a requirement of the universal approximation theorem that σ is neither linear nor a polynomial (Mhaskar *et al.*, 1992).

Neural networks composed of only affine operators are commonly referred to as 'Fully Connected Neural Networks' (FCNN) or 'Multi-Layer Perceptrons' (MLP). Neural networks composed of at least one convolution operator are referred to as 'Convolutional Neural Networks' (CNN). Other types of networks composed of a combination of affine, convolution, or other operators have also been proposed, such as transformers (Dosovitskiy *et al.*, 2020) and graph neural networks (Scarselli *et al.*, 2008).

2.2.2 Automatic differentiation

As introduced in Sec. 2.1.5, optimization algorithms most often require the computation of the gradient of the cost function L with respect to the optimized parameters θ, $\nabla_\theta L$ (Bonnans *et al.*, 2003b). To compute $\nabla_\theta L$, it is possible, if L has a closed-form expression, to manually derive and code the exact expression of $\nabla_\theta L$. This approach, however, is quite time-consuming and does not adapt to changes in the expression of L or g_θ. When it is not practical or possible to compute $\nabla_\theta L$ in this way, for example if L or g_θ do not have a closed-form expression, numerical differentiation can alternatively be employed.

Numerical differentiation is often implemented with finite differences, expressed as $\frac{\partial f}{\partial x}|_{x_0} = \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h} \approx \frac{f(x_0+h)-f(x_0)}{h}$ when h is small, with f a function of x and x_0 a point where f is evaluated. The issue with finite difference is the necessity of very small h to correctly approximate the gradient, especially in high dimensions, which in turn leads to a high number of iterations for the optimizer to converge. Additionally, two evaluations of f must be computed for each parameter that needs to be differentiated against, leading to high computational costs.

In order to always use the exact expression of $\nabla_\theta L$ without having to derive and code it explicitly, solving the above issues, automatic differentiation has been increasingly used in the last decade, being almost ubiquitous in the deep learning field (Baydin *et al.*, 2018). Automatic differentiation leverages the fact that, in many problems, $\nabla_\theta L$ does in fact have a closed-form expression, even if it is very long. In order to avoid having to code it manually, the chain rule of derivation $\frac{\partial g_{\theta_N} \circ \cdots \circ g_{\theta_1}}{x} = \frac{\partial g_{\theta_1}}{\partial x} \prod_{i=2}^{n} \frac{\partial g_{\theta_i}}{\partial g_{\theta_{i-1}}}$, is employed to compute $\nabla_\theta L$ automatically. Practically, in programming frameworks supporting automatic differentiation such as PyTorch (Paszke *et al.*, 2017), this means that for every differentiable function g programmed in the framework, its derivative g' is also programmed. Then, when computing $L(g_\theta, f, x)$, the output value ω_i of each g_{θ_i} is stored, and $\nabla_\theta L$ is obtained by successively evaluating each $\frac{\partial g_{\theta_i}}{\partial g_{\theta_{i-1}}} = g'_{\theta_i}(\omega_{i-1})$.

2.2.3 Statistical learning

In statistical learning methods, one wishes to minimize L over a dataset X, as opposed to classical optimization-based methods, which aim to minimize L over a single data sample. Using gradient descent to minimize L over the whole dataset would normally require computing its gradient on the whole dataset, i.e. $\nabla_\theta|_X L = \sum_{x\in X} \nabla_\theta L(g_\theta, f, x)$. With this formulation, one iteration of the optimization algorithm would take $\mathcal{O}(|X|)$ evaluations of L. Considering that, often, both $\mathcal{O}(|X|)$ and the number of iterations required for convergence are $> 10^3$, it is not practical to compute $\nabla_\theta|_X L$ in this way.

Rather, deep learning algorithms employ stochastic gradient descent (SGD) methods to minimize L over X. In SGD methods, $\nabla_\theta|_X L \approx \nabla_\theta|_B L$ where $B = \{x_k, \ldots, x_{k+b}\}$ is a randomly sampled batch of b samples in X. It has been demonstrated that, given some assumptions about L, the *expected* value of L after T iterations is bounded by $\frac{||\hat{\theta}||\rho}{\sqrt{T}}$ (Shalev-Shwartz *et al.*, 2014), with ρ the Lipschitz constant of L. In other words, SGD algorithms optimize the parameters θ of a deep neural network on a dataset in an efficient, but stochastic, way, whereas classical optimization algorithms optimize θ on a dataset in a deterministic, but expensive, way.

A very commonly used SGD algorithm is the Adam (Kingma *et al.*, 2014) algorithm, described in Alg. 3. Adam (from *adaptive moment estimation*) uses the

Algorithm 3 Adam (Adaptive Moment Estimation)

Require: η: Learning rate
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $L(\theta, \cdots)$: Objective function depending on the parameters θ
$\theta \leftarrow \theta_{init}$: Initial parameter vector
$m_0 \leftarrow 0$ (Initialize 1st moment vector)
$v_0 \leftarrow 0$ (Initialize 2nd moment vector)
$i \leftarrow 0$ (Initialize timestep)
while $i < N$ & $l >$ threshold **do**
    $i \leftarrow i + 1$
    $l \leftarrow L(\theta_{i-1}, \cdots)$
    $g_i \leftarrow \nabla_\theta l$ (Get gradients w.r.t. objective at timestep i)
    $m_i \leftarrow \beta_1 \cdot m_{i-1} + (1 - \beta_1) \cdot g_i$ (Update biased first moment estimate)
    $v_i \leftarrow \beta_2 \cdot v_{i-1} + (1 - \beta_2) \cdot g_i^2$ (Update biased second raw moment estimate)
    $\hat{m}_i \leftarrow m_i / (1 - \beta_1^i)$ (Compute bias-corrected first moment estimate)
    $\hat{v}_i \leftarrow v_i / (1 - \beta_2^i)$ (Compute bias-corrected second raw moment estimate)
    $\theta_i \leftarrow \theta_{i-1} - \eta \cdot \hat{m}_i / (\sqrt{\hat{v}_i} + \epsilon)$ (Update parameters)
end while
return θ_i (Resulting parameters)

first and second momentum of $\nabla_\theta L$ to update the *learning rate* η depending on the previous values of $\nabla_\theta L$. The learning rate of the optimizer is a key hyperparameter in learning-based problems. If η is too high, the optimization can fail entirely and if it is too small, the optimization can take a very long time. A common range for η is $10^{-3} > \eta > 10^{-6}$. (β_1, β_2) are also hyperparameters of Adam, although they are usually set to their default values, 0.9 and 0.999 respectively. Despite lacking proofs on its convergence, Adam is a widely used algorithm due to its strong empirical performances over other optimizers.

Additionally, in order to better constrain the values of θ and obtain more consistent neural network outputs, *regularization* is often employed. When using regularization, L is modified such that $L(\theta, \cdots) := L_{\text{data}}(\theta, \cdots) + \lambda L_{reg}(\theta)$ where λ is the regularization weight and L_{reg} is a cost function that depends only on θ and penalizes some characteristic of θ, for example its L_2 norm. Because vanilla L_2 regularization is not effective in Adam, AdamW, an alternative algorithm based on Adam, was proposed in (Loshchilov *et al.*, 2017) and widely used since, leading to improved performances. Another source of improvement is to use a learning rate scheduler to vary the value of η during the learning process. Usually, but not always, η is set to a relatively high value at the start of the learning, leading to a quick minimization of L, and then decayed to optimize θ more finely.

Both the regularization of θ (integrated into AdamW) and the decay of η require setting at least 1 additional hyperparameter each. With the multiplication of the number of hyperparameters, related to the optimization algorithm or the design of the network, finding optimal hyperparameter values has become a non-negligible problem in deep learning. A commonly employed, albeit very costly, method to find the best combination of hyperparameters is to run the learning algorithm again and again while varying the value of one hyperparameter at a time, until many combinations of hyperparameter values have been tested.

Thanks to statistical learning methods, extensive datasets and massively parallel computing, it has become possible to train neural networks with billions of parameters. This large number of parameters allows deep neural networks to represent very complicated transformations. Numerous, substantial, improvements have been brought about with the development of deep neural networks, notably in the field of image processing, which is one of the first domains revolutionized by deep neural networks. In the next section, a short overview of deep learning techniques developed for image processing and, specifically, for image registration, is proposed. This section focuses on deep learning methods for 3D-3D registration, since most deep learning-based registration methods in the literature have been developed for this application. A more complete review of deep learning-based fluoroscopy to CT deformable registration methods is available in Chap. 3.

2.2.4 Deep learning for 3D-3D registration

In the early 2010s, Convolutional Neural Networks revolutionized the field of image processing (Ciresan *et al.*, 2011; Krizhevsky *et al.*, 2017) with substantially improved performances, improving even upon previous deep learning approaches (Ranzato *et al.*, 2006). Since then, CNNs have been increasingly employed for medical image registration, offering drastically improved speed and achieving state-of-the-art performances in many cases (Fu *et al.*, 2020). Initially, CNN-based registration methods have been mostly applied to 3D-3D brain MRI

deformable registration (G. Wu *et al.*, 2015; Simonovsky *et al.*, 2016; Ghosal *et al.*, 2017).

Brain MRI deformable registration is an important area of research in registration, with applications to neurodegeneration quantification (B. B. Avants *et al.*, 2008) and brain shift compensation (Gerard *et al.*, 2017). On the IXI (I. Dataset, n.d.) and ADNI (A. Dataset, n.d.) MRI datasets, Ghosal *et al.* (Ghosal *et al.*, 2017) reported improvements over baseline demons algorithm (Thirion, 1998) between 4.3 and 13.6% in SSIM (Structural Similarity Index Measure) when combining the demons algorithm with a CNN. The fully CNN-based method in (Balakrishnan *et al.*, 2019) reported similar performances to the baseline SyN (B. B. Avants *et al.*, 2008) registration algorithm on a collection of datasets. However, the CNN-based method presents a 20,000x speed-up over SyN, thanks to its efficient GPU implementation and its non-iterative nature, with a registration time < 1 s, as opposed to 160 min for SyN. Finally, new deep learning approaches based on transformers (Dosovitskiy *et al.*, 2020), a novel kind of network architecture using linear layers, show promises to improve upon the baseline SyN registration method on both MRI and CT registration tasks (J. Chen *et al.*, 2022; Y. Tang *et al.*, 2022), while maintaining runtimes on the order of 1 s.

However, while deep learning-based registration methods have shown superiority to traditional methods, they are still scarcely used in practice due to a lack of tools adapted to clinical applications, a lack of large-scale clinical studies on their effectiveness and accuracy in clinical practice, and a lack of publicly available datasets for tasks other than brain MRI registration (X. Chen *et al.*, 2021). While deformable 3D-3D registration methods represent a valuable diagnostic tool, their use in an interventional setting is limited by 3D imaging devices requirements. Interventional 3D scans acquisitions necessitates interventional suites equipped with 3D imaging devices, can not be performed simultaneously with operational manipulation and, in the case of repeated CT scans, can expose the patient to important radiation doses. This motivates the development of 2D-3D registration methods, which are the focus of the following chapter.

Chapter 3

State of the art in 2D-3D deformable registration

Due to the previously mentioned drawbacks of 3D imaging, 2D imaging is often preferred in interventional settings. Optical modalities such as monoscopic, stereoscopic and endoscopic imaging, as well as ultrasound imaging and fluoroscopy are commonly used 2D imaging modalities for image guided interventions.

Compared to other modalities, fluoroscopic images offer a large field of view and fully show the internal anatomy of the patient, at the cost of a small but non-negligible radiation dose for the patient and the clinicians. Common fluoroscopy guided procedures include endovascular procedures, where fluoroscopy is used to follow a catheter in the vessels, percutaneous procedures where fluoroscopy is used to visualize a needle inserted through the skin and orthopedic procedures where fluoroscopy is used to image and operate on the bones (Rehani *et al.*, 2010).

An important issue with fluoroscopic images is the lack of contrast between tissues of similar density. This lack of contrast prevents clinicians from clearly distinguishing anatomical structures in the fluoroscopic image. In orthopedic surgery, this issue is mitigated by the important difference in density between bones and surrounding tissues.

In other types of interventions, clinicians often rely on contrast agent injection to increase the contrast between anatomical structures of interest and surrounding tissues. However, this solution introduces its own set of complications. Due to their nephrotoxicity (McClennan, 1990), contrast agents may only be used in limited quantities during procedures. Furthermore, blood flow causes the contrast effect to dissipate rapidly, necessitating repeated injections to maintain visibility.

As an alternative, radio-opaque markers may be implanted near target structures, although at the cost of an additional procedure prior to the operation and potential complications.

A way to mitigate these issues is to increase the information displayed in the

fluoroscopic images. This is the goal of 2D-3D deformable registration methods, wherein a preoperative, information-rich, 3D image is registered to an intraoperative, real-time, 2D image. Since preoperative CT scans are routinely acquired in many interventions and share the same image formation process as fluoroscopic images, fluoroscopy to CT registration naturally emerge as a solution to enhance fluoroscopy-guided interventions.

An alternative approach to registration is direct structure localization. While such approaches prove valuable in procedures like diagnostic angiography, where no preoperative CT images are available, it may not suit interventions that would benefit from preoperative information visualization.

The next sections will present state-of-the-art methods aiming to enhance fluoroscopy-guided interventions. This literature review will most notably focus on 2D-3D fluoroscopy to CT deformable registration methods, since these methods are closest to this thesis work.

First, section 3.1 presents 2D localization methods for fluoroscopy-guided interventions. Section 3.2 then provides an overview of 2D-3D rigid registration methods, a necessary first step in the registration process. A thorough review of state-of-the-art 2D-3D deformable registration methods is presented in section 3.3. Finally, section 3.4 targets biomechanical model-based registration methods, which could be used in fluoroscopy to CT registration to better handle the insufficient information content of fluoroscopic images.

3.1 2D localization

3.1.1 2D marker-based localization

Commercial systems such as the CyberKnife®(Adler *et al.*, 1997) have been developed to perform marker-based 3D tumor tracking. This system uses biplane orthogonal fluoroscopic image acquisitions to track, in 3D, the position of a marker implanted near the tumor. The error of this tracking system was found to be < 4 mm during 95% of tracking in a liver radiotherapy study (Winter *et al.*, 2015). However, using markers to track tumors is not an optimal solution, since marker implantation requires an additional invasive procedure and can be unreliable due to possible migration of the markers (Kitamura *et al.*, 2002). Nevertheless, systems like the CyberKnife® remain widely employed due to a lack of alternative, clinically approved products.

3.1.2 2D markerless localization

The Bayesian approach presented in (Shieh *et al.*, 2017) tackles the problem of 3D markerless tumor localization using a combination of an extended Kalman filter and a template matching algorithm. The method works on a time series of CBCT projections, producing a 3D tumor position distribution $\boldsymbol{P}_i$ for each image i in the series, computed using the previous distribution $\boldsymbol{P}_{i-1}$ and a patient-specific respiratory motion model. This approach achieved a mean 3D error comprised between 1.6 and 2.9 mm in a retrospective study on 13 cases. The error was measured using implanted markers as reference, which were removed from the image for inference. The computation time was greater than 1 second, which is not compatible with real-time applications. While the accuracy of this method is promising, it relies on a 3D respiratory motion model computed using the full CBCT time series, making this method unsuitable for clinical use.

In (Hirai *et al.*, 2019), Hirai *et al.* developed a Neural Network based markerless tumor localization method from fluoroscopic images. The training data was obtained from Digitally Rendered Radiographs (DRRs) generated from planning 4D-CT augmented by rigid translations and deformations at the global and local scale. The network input is a set of sub-images cropped from the input fluoroscopic image, and its output is a 2D target probability map. The weighted average of the target probability map is the predicted tumor position in 2D. The 3D tumor position is then obtained using the 2D prediction of the network on two orthogonal input images. The network is an encoder-decoder CNN with 25 layers. The accuracy achieved by this method was 2.18 ± 0.89 mm (3D euclidean distance) for a computation time of 39.8 ± 3.7 ms. These results show that, through the use of two orthogonal fluoroscopic images, markerless 3D tumor localization in real-time is possible.

In (W. Zhao *et al.*, 2019), W. Zhao *et al.* used a modified version of the VGG16 network (Simonyan *et al.*, 2014) (named after the VGG research team) to localize a tumor in fluoroscopic images. First, features are extracted from three fluoroscopic images, acquired at two orthogonal and an oblique incidence are acquired. Then, an additional three-layer CNN is employed to generate region proposals and their respective score from the feature maps. Finally, the features corresponding to these region proposals are input to five successive fully connected layers to obtain the planning target bounding box on the fluoroscopic image. The network is trained and tested on the same patient specific 4DCT, which does not allow to evaluate its performances in a realistic clinical setting. On synthetic data, the authors report a mean error < 2.6 mm for a computation time between 100 and 200 ms. While this level of accuracy is comparable with marker-based methods, this method is not real-time.

The method in (Zhang, X. Huang, Wang, *et al.*, 2020) uses a U-Net architec-

ture combined with intensity-based registration and patient-specific biomechanical modeling to accurately localize liver tumors. The number of projections required, 20, requires non-standard intraoperative imaging equipment, thus limiting the clinical applicability of the method.

In (Y. Yan *et al.*, 2024), Y. Yan *et al.* developed a method to perform lung tumor tracking from dual plane color fluoroscopic acquisition. The pair of color fluoroscopic images are first converted to grayscale fluoroscopic images using a style transfer U-net. Then, the images are further processed by another U-net to suppress bones. Finally, a third U-net is used to detect the tumor in images. The authors train and test on a dataset of preoperative 4D-CT data and intraoperative color fluoroscopic images for the style transfer network and report accuracies between 0.41 and 1.1 mm for 7 patients and from 6 to 10 mm for the other 3 patients. In these last 3 cases, the cause of failure was identified by the author to be the partial overlap between the tumor and the liver in the fluoroscopic images. The accuracy of the method is evaluated by comparing the predicted tumor region and the manually annotated tumor region.

While localization methods demonstrate potential in radiotherapy applications, their utility may be reduced in other procedures. For example, in lung nodule resection procedures, pneumothorax renders fine anatomical structures invisible in fluoroscopic and sometimes CBCT images (Rouzé, 2022) In percutaneous procedures, surgically induced deformations might also reduce the accuracy of these methods, trained solely on respiratory deformations. Finally, these methods do not merge preoperative data with intraoperative images, as opposed to 2D-3D registration methods that are presented in Sec. 3.2 and Sec. 3.3.

3.2 2D-3D rigid registration

As introduced in Sec. 2.1.2, the first step in image registration is rigid registration. In 2D-3D rigid registration, one wants to compute a 3D affine + 3D-2D projective transformation that aligns a 3D volume to a 2D image. Seminal works used intensity-based optimization methods, such as BFGS, Covariance Matrix Adaptation Evolution Strategy (CMA-ES) or Bound Optimization BY Quadratic Approximation (BOBYQA), to solve this problem (D. C. Liu *et al.*, 1989; Berger *et al.*, 2016; Hansen *et al.*, 2003; Powell *et al.*, 2009). Other solutions are feature-based methods, that rely on the detection of either anatomical (Wunsch *et al.*, 1996; Benameur *et al.*, 2003) or artificially implanted (Gall *et al.*, 1993; T. S. Tang *et al.*, 2000) landmarks.

One of the first, if not the first, deep learning method for rigid 2D-3D registration was proposed in 2012 (Gouveia *et al.*, 2012). It uses 6 separate MLPs to model the 6 Degrees Of Freedom (DOF) of the rigid transformation from a fluoro-

scopic image. The input fluoroscopic image is first pre-processed to extract global image features, which are then input to the 6 MLPS to compute the rigid transform. It is validated on synthetic digital subtraction angiography images, which are contrast-enhanced fluoroscopic images processed to only show contrast-enhanced vessels without the surrounding anatomy. Recently, CNN-based methods have been tested on real neurological digital subtraction angiography images, with an average error ranging from 5.8 to 6.3 mm (D.-X. Huang *et al.*, 2024) between projected vessel positions and ground truth (manually annotated) positions. 2D-3D rigid registration methods have also been validated on real fluoroscopic images of the hip (Jaganathan *et al.*, 2023; Gao, Killeen, *et al.*, 2023; Gao, Feng, *et al.*, 2023; Gopalakrishnan, Dey, *et al.*, 2024) and the spine (M. Chen, Z. Zhang, Gu, Ge, *et al.*, 2024; M. Chen, Z. Zhang, Gu, and Kong, 2024). Finally, a rigid registration approach robust to non-rigid motion has also been developed and validated on real fluoroscopic images, with a mean accuracy of 14.1 mm in 2D Projection Distance (PD) (B. C. Lee *et al.*, 2022).

These advancements show the potential of such methods for robust pose recovery as a first step in the registration pipeline, before applying deformable registration to recover organs' deformations.

3.3 2D-3D deformable registration

A common objective of many existing fluoroscopy-to-CT deformable registration methods is the compensation of either respiratory or cardiac motion in fluoroscopy-guided interventions. Consequently, these 2D-3D deformable registration use a Statistical Deformation Model (SDM) as an inductive bias to recover deformations, contrary to 3D-3D registration methods that most commonly do not employ an SDM to model deformations. Another point common to the majority of 2D-3D deformable registration methods in the literature is that they are often not validated on real fluoroscopic images, but rather on synthetic fluoroscopic images generated from a CT volume. This is due to the fact that the 'ground truth' registration transform can not be known when registering a (moving) CT to a fluoroscopic image, except in the case where the fluoroscopic image is paired with another (fixed) CT, in which case it is possible to compare the moving CT after registration to the fixed CT.

3.3.1 Breathing SDM-based registration methods

In the following methods, a cyclical breathing motion pattern is extracted using registration from a preoperative 4DCT. Then, a PCA is computed from the DVFs representing the breathing motion, and the first components that explain more

than $\simeq 95$ % of the variance are kept to form build a breathing motion SDM, as detailed in 2.1.5.

In (C.-R. Chou, Frederick, *et al.*, 2013), linear operators are optimized at test time to minimize the difference between the 2D projection of the deformed volume and the 2D image. First, an SDM is built from the preoperative 4DCT, keeping the first 3 components of the PCA. Then, at test time, the parameter estimation is performed in 4 iterations: at each iteration, K parameters of the SDM are sampled, for which the corresponding deformations and then projections of the preoperative CT are computed. Then, the difference between each projection and the input image is computed, and a linear regression is performed to find the combination of parameters that minimize the difference. Finally, the CT is transformed by the parameters found in the previous iteration and the process is repeated. The method is tested on experimentally acquired CBCT scans of 5 lung radiotherapy patients, which provide a 3D ground truth along with 2D fluoroscopic images. Before deformable registration, the CBCT are first registered to the preoperative CT. The authors report a mean Target Registration Error (mTRE) of 2.7 mm and 1 mm in 3D and 2D, respectively, for an average registration time of 2.6s on a 128-core Nvidia GeForce 9800 GTX GPU, potentially enabling real-time capabilities on modern hardware.

The method presented in (C.-R. Chou and Pizer, 2013) is similar to the previous method, except in the way the relationship between the SDM parameters and the fluoroscopic image is modeled. In this method, both a decision forest and a linear regression are optimized on a large number of samples at training time. The decision forest is a classification model, used here to estimate the most probable SDM parameters from a fluoroscopic image. A linear regression is then computed between image features and SDM parameters. Then, at test time, the input fluoroscopic image is first classified using the decision forest, and the associated SDM parameters are computed using the linear regression. The method was validated in the same way as in (C.-R. Chou, Frederick, *et al.*, 2013), with a similar mTRE and a faster runtime of $\simeq 70$ ms, making it suitable for real-time applications.

In contrast with previous methods, (M. D. Foote *et al.*, 2019) is one of the first method to employ deep learning for 2D-3D deformable registration. In this method, a 'rank-constrained diffeomorphic density matching' algorithm (detailed in (M. Foote *et al.*, 2017)) is used to extract the breathing motion from a preoperative 4DCT. DRRs are then generated by sampling the PCA components and projecting the deformed CT volume to form a training dataset. A DenseNet (Iandola *et al.*, 2014) CNN is then trained on this dataset to predict the value of the PCA components from the input DRR. Unfortunately, the model is validated on DRRs generated from the same 4DCT that was used to compute the PCA, so the performance of the method on unseen data is unknown.

In (Nakao *et al.*, 2022), Nakao *et al.*, proposed a deep learning framework to predict the deformation of abdominal organ meshes from a fluoroscopic image. Different from other methods, the displacement is only predicted at the surface of the organs and the displacement inside the organs remains unknown, which could hinder the capability of this method to predict the position of small structures such as tumors. In this method, a CNN is used to map an input fluoroscopic image to a 2D projection of the 3D DVF that would register the preoperative anatomy to the intraoperative anatomy. Then, a Graph Convolutional Network (GCN) is used to compute the 3D displacement for each vertex of the preoperative organ meshes from the projected DVF. The network is trained and tested on a DRR dataset generated from 124 3DCTs and 35 4DCTs, each manually annotated. The 4DCT scans are used to build a statistical atlas of deformation, from which a training dataset is built by varying the value of the first two PCA components. Notably, 12 4DCT scans were kept out of the PCA to test the method on unseen deformations. Additionally, the method was tested with random pose changes (translations of up to 17 mm but no rotations) and found to be robust to uncertainty in the pose. The method is able to predict abdominal organ shapes, with an average accuracy in 3D ranging from 3.5 mm to 6.1 mm at the organ surface, from an average initial displacement of 5.2 to 9.4 mm, depending on the organs.

The method presented in (Shao, Jing Wang, *et al.*, 2022) combines statistical motion recovery similar to above methods with a Finite Element Method (FEM) to perform 2D-3D registration of the liver. Similarly to the previous method, a network formed by the combination of a CNN and a GCN is trained on a PCA-generated DRR dataset to predict the displacement at the surface of the liver. Then, the surface displacement of the liver mesh is used as a boundary condition for the FEM which models the liver as a hyperelastic solid and computes the internal liver deformation by iteratively minimizing the stress. The authors report an average 3D tumor tracking accuracy of 1.1 mm, from an average initial displacement of 6.1 mm. While the FEM is a physically accurate way of recovering deformations, it suffers from a slower runtime, that the authors did not report, compared to purely deep learning based approaches, making it unsuitable for real-time applications. Furthermore, as in (M. D. Foote *et al.*, 2019), the method was validated on the same 4DCT that was used for training, so its performances on unseen data remain unknown.

Finally, in a recent work (Wijesinghe, 2024), Wijesinghe presented two methods to recover volumetric organ mesh deformation from a fluoroscopic image. In a variant, a CNN is used to extract features from the input fluoroscopic image, which are then passed to a Graph Neural Network (GNN) to predict the deformation of the preoperative liver mesh. In the second variant, an MLP autoencoder is trained to encode the organ mesh deformation into a latent space and then decode

it with minimal error. Then, a CNN containing a self-attention layer extracts features from the image, which are then input to another MLP to predict the latent representation of the organ deformation, to be compared against the latent representation obtained via the now frozen autoencoder, considered here as ground truth. At test time, the CNN and the two MLPs are used to obtain first the latent representation and then the actual deformation from the input image. Both methods employ a separately trained CycleGAN (J.-Y. Zhu *et al.*, 2017) network to perform style transfer between synthetic DRR images and real fluoroscopic images to improve similarity between synthetic training data and real testing data, as a way to bridge the domain gap. Additionally, the pose was varied during training by rotating the camera around the superior-inferior axis in a 360° arc. While the method is trained and tested on synthetic data generated with a PCA from a 4DCT, as in previous methods, reporting mTRE values below 1 mm, it was also qualitatively validated on real fluoroscopic images, showing a good agreement between projected organ shape and image content. Unfortunately, quantitative results on real images were not available.

3.4 Biomechanical model-based registration methods

The methods presented so far, with the notable exception of (Shao, Jing Wang, *et al.*, 2022), have parameterized the registering transform as a purely geometrical transform. However, these approaches suffer from a lack of realism because, in the real world, deformations are subject to physical constraints such as resistance to compression, stretching, tearing and the absence of self-intersections. More generally, physical laws impose heavy constraints on the type of deformations that are possible for deformable solids like internal organs. In order to respect these constraints, biomechanical model-based registration methods have been developed. These methods follow the general principles detailed in Sec. 2.1.3. In the next paragraphs, a few such methods are presented.

In (Broit, 1981), Broit developed the first biomechanical model-based registration method, where brain images are modeled using a linear elasticity model (see Eqn. 2.2), and forces are optimized to perform registration. This work introduces the use of anatomy atlases for registration, by first computing a mean anatomy is computed on an image dataset and then registering every subject in the atlas to the mean anatomy.

In (Rabbitt *et al.*, 1995), Rabbitt *et al.* presented one of the first methods to use a hyperelastic model for registration. The hyperelastic model is linearized by its Gâteau-Taylor series term and the FEM is used to compute the energy of the

deformation on a regular grid around the object. Newton's method is then used to minimize the energy of the deformation between the two images to be registered.

In (Pennec *et al.*, 2005), a new, invertible, variation of the St. Venant-Kirchoff elastic energy, dubbed Riemannian Elasticity, is proposed. In the St. Venant-Kirchoff model, the elastic energy depends on the Euclidean distance between the strain tensor of the deformation and the identity. The authors propose to replace the Euclidean distance with the Riemannian distance, which is the squared distance between the logarithm of the strain tensor and the logarithm of the identity (which is equal to zero). The authors show that with this new formulation, the deformation between a rest state and a deformed state has the same energy as the inverse deformation, a desirable property in elasticity. Additionally, the authors compute the mean and covariance matrix of the logarithmic strain on a population of organs to build a statistical Riemannian elasticity energy, a less computationally intensive operation in this new formulation. Unfortunately, the authors did not validate this statistical method on data. Instead, they showed that registration performances on a brain MRI dataset are equivalent between Riemannian elasticity and standard St. Venant-Kirchoff elasticity, at the cost of a three-fold increase in computation time for Riemannian elasticity.

In order to remediate to the slow computation of Riemannian elasticity while keeping the invertibility of the registering transform, Yanovski *et al.* (Yanovsky *et al.*, 2008) propose another variation of the St. Venant-Kirchoff energy. Starting from the St. Venant-Kirchoff energy as a function of the strain matrix, itself a function of the product between the displacement and its Jacobian, the authors note that the direct minimization of the energy is computationally expensive. Instead, they replace the displacement-Jacobian product in the energy by a new, unknown variable V to be optimized, with a penalty on the distance between V and the displacement-Jacobian product. In the paper, the authors only provide qualitative results on a pair of brain MRI images, the quantitative accuracy remaining unknown.

Biomechanical model-based methods have also been designed for specific clinical applications such as Video-Assisted Thoracoscopic Surgery (VATS) procedures. In VATS procedures, the intervention is performed by the clinician through surgical ports created in the abdominal cavity. These ports break the pressure equilibrium in the intrapleural space, leading to lung collapse, a phenomenon known as pneumothorax. Due to pneumothorax and change of pose of the patient from supine in the preoperative scan to lateral decubitus during the intervention, the localization of the nodule becomes unknown. In current surgical practice, this issue is resolved by intraoperative localization of the nodule using an interventional CBCT scanner (Rouzé, 2022). However, manual nodule localization using a CBCT scanner is a time-consuming process, that is prone to failure due to poor visibility of some

lung nodules. While this issue can be mitigated by preoperative fiducial implantation, it is at the cost of additional intervention time, radiation exposure and invasiveness. Another solution for intraoperative nodule localization is registration between the preoperative CT scan (containing the nodule segmentation) and the intraoperative CBCT scan.

In a seminal work (P. Alvarez, Rouzé, *et al.*, 2021), P. Alvarez, Rouzé, *et al.* proposed a biomechanical model-based registration method tailored to this application. The authors used a poroelastic constitutive law to model the lung in the context of deformations as extreme as pneumothorax. In this work, the deformation between the preoperative CT and intraoperative CBCTs is modeled in two parts. First, the deformation induced by the pose change between the CT and the first CBCT (acquired before pneumothorax) is recovered using an intensity-based method, and refined using the biomechanical lung model. Specifically, in the intensity-based registration part, rigid registration is first performed by segmenting the intraoperative spine and registering it with the preoperative spine. Then, the parameters of a B-spline deformation model are optimized with the Normalized Cross Correlation (NCC) similarity metric between intraoperative and preoperative scans. Finally, this initial intensity-based deformation is applied as a Dirichlet boundary condition on the surface of the lung finite element mesh and the internal deformation is computed using the FEM poroelastic model. The second part of the method handles the registration of the CBCT acquired after pneumothorax. As in the first part, the two images are first rigidly registered. Then, the parameters of a novel pneumothorax model are optimized to register the lung before and after pneumothorax. In this model, the pneumothorax effect is simulated by applying a hydrostatic pressure as a Dirichlet boundary condition on the lung surface. In the simulation, this causes the lung to shrink due to the evacuation of fluid in the poroelastic model. Additionally, a prior elastic registration is performed between the airways before and after pneumothorax. The resulting displacement is applied as Dirichlet boundary conditions to corresponding nodes in the lung model. Finally, a gravitational load is applied on the whole model. The parameters that are optimized to perform the registration are the material properties of the lung and the displacement of the diaphragm acting on the lung. The authors validate the method on 5 cases, and used anatomical landmarks to measure the registration accuracy. After rigid registration, the average pose change deformation was measured to be between 6.8 and 25.8 mm, depending on the case. After deformable registration of the pose change, the mean TRE was between 1.0 and 2.7 mm. For the pneumothorax registration, the initial average deformation was between 19.5 and 37.7 mm and the mean TRE after deformable registration was between 4.9 and 14.3 mm. These results show that biomechanical model-based registration in the context of large deformations and solid-fluid interactions is a very challenging

problem, that this seminal work managed to partially solve thanks to a multi-step approach.

In a related work (Lesage *et al.*, 2020), Lesage *et al.*, modeled the lung as a hyperelastic Ogden material, with a range of parameters obtained from a previous study on the mechanical characteristics of porcine lungs. The method performs registration between two CT scans, before and after pneumothorax. First, the airways, bronchi and thoracic cavity were segmented in both CT scans and registered using the Morfeus linear elasticity biomechanical model (Brock *et al.*, 2005). Then, these registered structures were added to the lung model and their position were used a Dirichlet boundary conditions, partially constraining the shape of the lung. Finally, the lung model was subject to gravity and the nodes at the top of the lung were subject to pressure to simulate the pneumothorax effect, with nodes corresponding to the airways outside the lung set to fixed positions, and the simulation was set to run until sufficient deflation was obtained. The authors varied the material parameters and the applied pressure for each patient, and reported the best mean TREs, which range from 6.0 to 16.0 mm. When using the best overall set of parameters for all patients, the accuracy ranged from 8.0 to 9.3 mm, excluding one patient.

Chapter 4

Domain-agnostic 2D-3D deformable registration

4.1 Introduction

As previously mentioned, the existing literature on 2D-3D deformable registration focuses on motion management for image-guided radiotherapy. In this context, a planning 4D CT image of the patient is often available, allowing to derive a respiratory motion model to train a neural network on DRRs exhibiting respiratory motion. Additionally, with some exceptions, methods in the literature are not validated on real fluoroscopic images due to the difficulty of obtaining ground truth.

In contrast with the existing literature, we aim to develop a 2D-3D fluoroscopy to CT registration method suitable for all fluoroscopy-guided interventions. In this context, the preoperative image is a 3D CT volume and the intraoperative deformation to recover is a combination of anatomical (i.e. respiratory) and intervention-related deformations. Consequently, unlike state-of-the-art 2D-3D fluoroscopy to CT registration methods, we cannot rely on a patient-specific respiratory motion model to recover anatomical deformations during the intervention.

Instead, we draw inspiration from another line of work (Shen *et al.*, 2019; Yikun Zhang *et al.*, 2021; J. Guo *et al.*, 2024) focusing on volume reconstruction from a few fluoroscopic projections. Notably, in (Shen *et al.*, 2019) Shen *et al.* showed that a single fluoroscopic image was sufficient to reconstruct a 3D volume of the patient with an acceptable accuracy. In this work, an encoder extracts 2D features from an input fluoroscopic image, which are then transformed into 3D features and then decoded to output a reconstructed CT volume of the patient. Specifically, the encoder is a CNN based on the ResNet architecture (K. He *et al.*, 2016), which groups convolutional layers into residual blocks. A residual block is

a sequence of N ($N = 1$ in (Shen *et al.*, 2019)) convolutional layers processing the input sequentially, with a residual connection to add the input of the block to the output of the block. This operation has been shown to improve performances and mitigate the vanishing gradient problem (K. He *et al.*, 2016; Szegedy *et al.*, 2017; F. He *et al.*, 2020). In this work, the kernel size and stride parameters of the encoder are set up such that every two layers in the encoder doubles the number of channels while halving the spatial resolution, resulting in a compression of the data by a factor of two. Every other layer in the encoder keeps the number of feature and the spatial resolution constant. The authors performed an ablation study to determine the optimal number of layers in the encoder and found that 1+10 layers offered the best performances, with the first layer transforming the input grayscale image into a 256-channel, half resolution feature map.

Following the encoder, a transformation module is used to transform 2D features into 3D features. This module applies a reshape operation in which the number of channels is divided by two to form an additional spatial dimension. Finally, the decoder, a 3D CNN which does not use skip connections, upsamples the 3D features to output a high resolution predicted CT volume. Inversely to the encoder, the decoder doubles the spatial resolution and halves the number of channels at every other layer. The key point of this architecture is the direct transformation of 2D features into 3D features, which enables the prediction of a 3D image from a 2D input.

The authors demonstrate that this 2D-3D architecture is able to reconstruct a 3D volume of the patient from a single fluoroscopic image, using a synthetically generated patient-specific dataset composed of pairs of CT volumes and corresponding DRRs of the patient. The training dataset is generated by translating, rotating, and deforming the original CT volume of the patient and rendering DRRs images from the transformed CT volumes. Unfortunately, the authors did not elaborate on the process used to produced deformed CT volumes. While the predicted 3D volumes resemble ground truth volumes, fine anatomical structures are either not recovered or misshapen. Nonetheless, this work serves as a proof of concept for our application, demonstrating that dense 3D information can be recovered from a single-view 2D fluoroscopic image.

Inspired by the success of this approach, and its relevance to our application, we build our 2D-3D registration network upon this architecture, with incremental improvements to raise the accuracy of the method. As in (Shen *et al.*, 2019), our architecture is also split into an encoder, transformation module and decoder part. However, instead of reconstructing a 3D volume from the fluoroscopic image, our goal is to register preoperative 3D CT data to intraoperative fluoroscopic images. To achieve this, we modify the last layer of the decoder to output a 3D Deformation Vector Field (DVF) that registers the CT volume to the fluoroscopic image.

The 3D DVF is represented as a 3-channel 3D volume, with each channel representing the displacement of voxels along one of the three anatomical axes. Using a discrete DVF representation allows for arbitrary deformations, even unfeasible ones presenting self-intersections. This is why, in traditional, iterative registration methods, such as (Thirion, 1998; Trouve *et al.*, 2005; B. B. Avants *et al.*, 2008), parametric representations are used. However, more recent, learning-based, single-shot registration methods using fully convolutional networks or transformer networks, have shown that registration could be performed efficiently, with limited self-intersections, using a discrete representation of deformations (De Vos *et al.*, 2017; Shan *et al.*, 2017; Miao *et al.*, 2018; Balakrishnan *et al.*, 2019; J. Chen *et al.*, 2022; Y. Zhu *et al.*, 2022). Thus, given the real-time requirements of our application, we adopt a discrete representation of deformations as well.

These single-shot registration methods are often trained in a self-supervised manner, using pairs of clinically acquired 3D images. The network predicts a DVF to register a 'moving' 3D image to a 'fixed' 3D image, and an image intensity-based loss function is used to optimize its parameters. In our context, the goal is instead to register a moving 3D image to the anatomy observed in a fixed 2D image, which is an ill-posed problem, as multiple 3D images can correspond to a single 2D projection. Due to this ill-posedness, training a 2D-3D deformable registration network on pairs of clinically-acquired 2D and 3D images may not be feasible.

However, deep learning is not constrained by the availability of real training data. In fact, it is a relatively common approach to generate large amounts of synthetic data to train a neural network (Hoffmann *et al.*, 2021; Tobin *et al.*, 2017; Dahmen *et al.*, 2019; Hu *et al.*, 2023; Doersch *et al.*, 2019) to perform tasks where few or no labeled data are available. This strategy is employed in much of the literature on deep learning-based 2D-3D registration, where synthetic 2D projections are generated from clinically acquired 3D volumes to create a paired 2D-3D synthetic dataset. From there, the learning objectives vary between target localization, DVF estimation, or volume reconstruction (see Sec. 3.1, and 3.3 for an overview of such methods).

In this chapter, our baseline 2D-3D deformable registration method is presented and evaluated through several (published and unpublished) works. Sec. 4.2 details the domain agnostic data generation method and neural network architecture, with three key contributions. First, we propose a novel registration method able to recover deformations commonly encountered in fluoroscopy-guided interventions, moving beyond the limitations of existing methods focused on respiratory motion. In this aspect, our main contribution is the domain agnostic data generation method (Sec. 4.2.2), which eliminates the need for prior knowledge of motion during the procedure, making the registration agnostic to the domain of defor-

mations. Second, we developed and evaluated a novel backprojection module for 2D-3D registration that transforms 2D features into 3D features while considering projective geometry (Sec. 4.2.3.1). Third, we propose a novel loss combination (Sec. 4.2.3.2), specifically designed for 2D-3D deformable registration, where the network is supervised in projective space rather than anatomical space, leading to improved performances.

Then, Sec. 4.3 presents an initial study on 2D-3D deformable registration that compares our domain agnostic data generation method with the traditional PCA-based data generation method employed in the literature.

In Sec. 4.4, a follow-up study demonstrating the potential of the method to reduce the need for contrast agent injection in percutaneous interventions is presented. The presented data generation method and neural network are similar to the work in Sec. 4.3 but the clinical application and experimental results differ. This study was published in the proceedings of the 2023 Hamlyn Symposium on Medical Robotics (HSMR):

> Francois Lecomte *et al.* (2023). "Enhancing fluoroscopy-guided interventions: a neural network to predict vessel deformation without contrast agents". In: *The Hamlyn Symposium on Medical Robotics*. The Hamlyn Centre, Imperial College London London, UK, pp. 75–76

In Sec. 4.5, a study submitted for publication (November 2023) in the Medical Image Analysis (MedIA) journal is partially reproduced. This study utilizes the most recent version of our method, presented in Sec. 4.2. Compared to previous works, we perform a more thorough experimental validation of our method. Notably, it is evaluated for the first time on real, surgically induced deformations of porcine animal models. This work also includes a comparison between our method and a state-of-the-art 2D-3D deformable registration on breathing motion prediction and intervention-related deformation prediction.

Finally, in Sec. 4.6, our method is used in combination with a deep reinforcement learning algorithm, developed by V. Scarponi *et al.*, for autonomous catheter navigation. The deep reinforcement learning navigation algorithm and our 2D-3D registration network are trained separately and are combined during evaluation, with the registration network being used to update the vessel model in real-time during navigation. In the study, synthetic results show that autonomous navigation is possible in the liver and heart, with improved success rates when using our registration method to update the vessel model during navigation. This study, co-authored with V. Scarponi *et al.*, was presented at the 2024 International Conference on Intelligent Robots and Systems (IROS):

> Valentina Scarponi, François Lecomte, *et al.* (Oct. 2024). "Autonomous Guidewire Navigation in Dynamic Environments". In: *2024*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)

4.2 Deep learning framework

The contents of this section are reproduced from the 'Method' section of the previously mentioned study to be published in the MedIA journal. In each study, differences with the method presented below are clearly indicated. This section is structured as follows: Sec. 4.2.1 provides an overview of the method and clinical workflow, Sec. 4.2.2 presents our synthetic data generation process, Sec. 4.2.3 presents our deep learning approach, and Sec. 4.2.4 presents the data augmentation technique used to better handle the domain gap between synthetic and real fluoroscopic images. An additional section, not presented in the MedIA study, Sec. 4.2.5, details the pre- and post-processing steps applied to the input and output of the network.

4.2.1 Overview

Our framework is based upon the most common steps of fluoroscopy-guided interventions, outlined in Fig. 4.1.

First, a preoperative CT scan is acquired and structures of interest are segmented. The intervention can then be planned by the clinicians. We assume that the pose of the C-arm with respect to the patient is determined during this step. During the intervention, the C-arm is positioned as per planning and is used for intra-operative image guidance.

In the context of image-guided liver therapies, contrast agents are injected before acquiring the preoperative CT. From the preoperative Contrast Enhanced CT scan (CECT), the vessel tree is segmented to obtain a 3D segmentation volume. Using the CECT and the planned C-arm pose, we generate a synthetic training dataset composed of synthetic deformations and synthetic fluoroscopic images, as described in Sec.4.2.2 and as shown in Fig. 4.1..

The neural network is then trained on the synthetic dataset (see Sec. 4.2.3). Next, during the intervention, a fluoroscopic image of the patient is acquired without contrast agents. The fluoroscopic image is processed in real time by the network to compute the deformation between the preoperative CT scan and the intraoperative anatomy. The deformation computed by the network is used to warp the preoperative vascular tree. Finally, the warped tree can be projected on top of the fluoroscopic image (i.e., augmented fluoroscopy) using ray-casting techniques, to obtain a result similar to the image in the lower right part of Fig. 4.1.

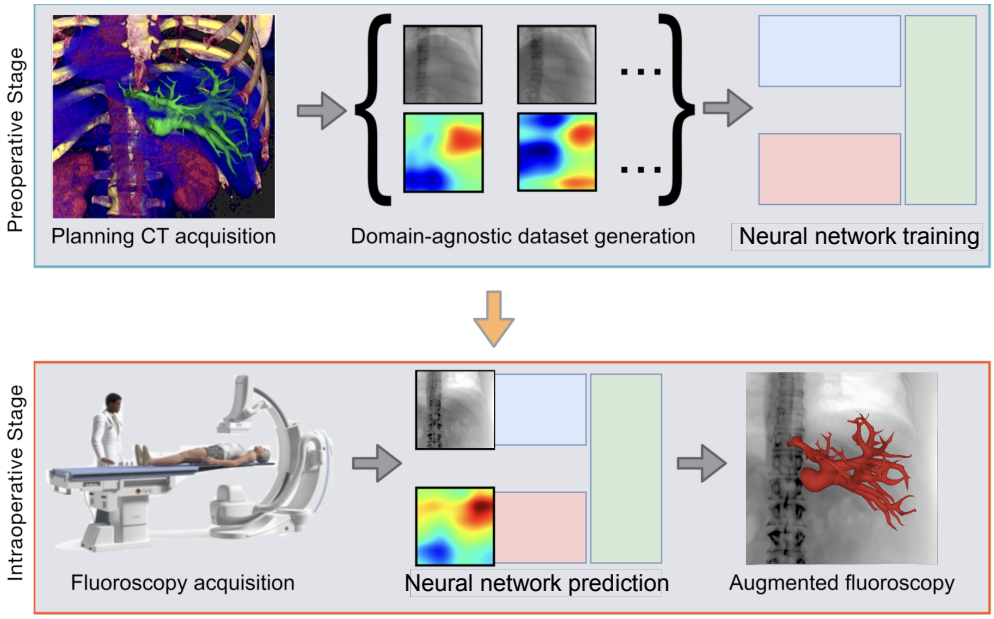


Figure 4.1: Overview of the proposed method. First, the intervention is planned from a 3D CT scan of the patient, where structures of interest are segmented, and the C-arm pose is determined (top left). Second, the neural network is trained on non-rigid deformations of the 3D CT and synthetic fluoroscopic images (top, middle, and right). Here, non-rigid deformations are represented schematically in 2D (in reality, deformations are 3D vector fields), with color to indicate the amplitude of the displacement. Third, during the intervention, the C-arm is positioned, and a fluoroscopic image is acquired (bottom left). Fourth, the network computes the deformation from the fluoroscopic image (bottom middle) and the warped vessel tree is used to augment the fluoroscopy (bottom right).

4.2.2 Data generation

To train the network to estimate a deformation $\boldsymbol{\phi}$ from a non-contrasted fluoroscopy p, we generate a training dataset composed of pairs of synthetic $\boldsymbol{\phi_i}$ and p_i. To generate synthetic non-contrasted fluoroscopic images, we first transform the preoperative Contrast Enhance CT image I_{CE} into a non-contrasted CT image I via inpainting ((Barnes *et al.*, 2009)). This operation preserves the information outside the segmentation and removes as much contrast information as possible. We can then generate deformed, non-contrasted CT images I'_i from I.

For each sample of the dataset, the process goes as follows: The deformation $\boldsymbol{\phi_i}$ is generated using a sum of randomized Gaussian kernels (Sec. 4.2.2.1). Then, the deformed CT image I'_i is generated from I and $\boldsymbol{\phi_i}$ (Sec. 4.2.2.2). Finally, following (4.4), we generate synthetic fluoroscopic images p_i from I'_i. (Sec. 4.2.2.3).

4.2.2.1 Deformation generation

A non-rigid deformation is defined on the 3D image $I(\mathbf{x})$ by $\boldsymbol{\phi}(\mathbf{x}) : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ with $\mathbf{x}$ a point in the image volume and $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x} + \boldsymbol{\varphi}(\mathbf{x})$, with $\boldsymbol{\varphi}(\mathbf{x})$ a displacement vector field. We restrict the region where $\boldsymbol{\phi}$ is defined to a sub-region of the volume, which will be referred to as the field domain. The key characteristics we seek in the displacement field are smoothness and invertibility. A good candidate for producing such displacement fields is the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework ((Trouve *et al.*, 2005)), which demonstrated very good performance in non-rigid registration problems ((Durrleman *et al.*, 2014)). In this framework, the non-rigid deformation $\boldsymbol{\phi}$ that registers an image I to an image I' is obtained by integrating a velocity field $\mathbf{V}(t, \mathbf{x})$ over time, following a set of differential equations to drive the evolution of $\mathbf{V}(t, x)$.

The authors demonstrate that $\mathbf{V}(t, \mathbf{x})$ can be expressed as:

$$\mathbf{V}(t, \mathbf{x}) = \sum_{k=1}^{N_{cp}} \boldsymbol{\alpha_k}(\boldsymbol{t}) \cdot \mathbf{K_k}(\mathbf{x}, \mathbf{y_k}(\mathbf{t})) \tag{4.1}$$

where $\mathbf{K_k}(t)$ are elements of a Reproducing Kernel Hilbert Space, such as Gaussian kernels, located at the N_{cp} control points $\mathbf{y_k} \in \mathbb{R}^3$ and associated with weights $\boldsymbol{\alpha_k} \in \mathbb{R}^3$. The Displacement Vector Field (DVF) $\boldsymbol{\varphi}$ is then given by $\boldsymbol{\varphi}(\mathbf{x}) = \int_0^1 \mathbf{V}(t, \mathbf{x}) dt$.

In our framework, we instead directly compute the DVF $\boldsymbol{\varphi}(\mathbf{x})$ by randomizing the control points $\mathbf{y_k}$, covariance matrices $\boldsymbol{\sigma_k} \in \mathbb{R}^{3\times3}$ and weights $\boldsymbol{\alpha_k}$ of the Gaussian kernels.

$$\boldsymbol{\varphi}(\boldsymbol{x}) = \sum_{k=1}^{N_{cp}} \boldsymbol{\alpha_k} \cdot \mathbf{K}(\mathbf{x}, \mathbf{y_k}, \sigma_\mathbf{k}) \tag{4.2}$$

To sample $\mathbf{y_k}$, we generate a set of random points in the field domain. We discard points that are within a threshold distance Δ_y of each other to prevent sharp variations in $\boldsymbol{\varphi}$, and re-generate rejected points until the desired number of control points is obtained. $\boldsymbol{\alpha_k}$ are sampled from a 2D uniform distribution and then multiplied by a random common scaling factor between 0 and 1 to ensure samples with small overall displacements are represented in the dataset. Finally, $\boldsymbol{\sigma_k}$ is generated as $N_{cp} \times 3 \times 3$ i.i.d variables with values between 15% and 30% of the size of the field domain. While this new formulation does not preserve the diffeomorphic nature of $\boldsymbol{\phi}$ by construction, we can compute its spatial Jacobian J and verify that it is positive, and if not, regenerate $\boldsymbol{\phi}$, thus ensuring diffeomorphic deformations.

4.2.2.2 Image warping

As stated above, we model the non-rigid deformations of the anatomy as coordinate transforms $\boldsymbol{\phi}(\mathbf{x})$. The warped CT image $I'(\mathbf{x}) = I \circ \boldsymbol{\phi}(\mathbf{x})$ is obtained by linearly interpolating the values of I at $\mathbf{x}' = \boldsymbol{\phi}(\mathbf{x})$:

$$I'(\mathbf{x}) = \sum_{\mathbf{z} \in \mathcal{Z}(\mathbf{x}')} I'(\mathbf{z}) \prod_{d \in \{0,1,2\}} (1 - |\mathbf{x}'_{\mathbf{d}} - \mathbf{z}_{\mathbf{d}}|) \tag{4.3}$$

where $\mathbf{z}$ are the 8 voxels nearest to $\mathbf{x}'$, d iterating through the 3 spatial components of $\mathbf{x}'$ and $\mathbf{z}$, and $\mathbf{z}$, $\mathbf{x}$ and $\mathbf{x}'$ expressed in voxel coordinates. This operation is known as backward warping.

4.2.2.3 Digitally Reconstructed Radiographs

We generate Digitally Reconstructed Radiographs (DRR) using the DeepDRR framework ((Unberath *et al.*, 2018)).

This framework models the C-arm as a pinhole camera, parameterized by a projection matrix $\mathbf{P}$ composed of an intrinsic matrix $\mathbf{H} \in \mathbb{R}^{3\times3}$ and an extrinsic matrix $\mathbf{E} \in \mathbb{R}^{3\times4}$. $\mathbf{E}$ is obtained from the planned pose of the C-arm and $\mathbf{H}$ is obtained from the characteristics of the C-arm detector panel. In the DeepDRR framework, the fluoroscopic image p observed during the intervention is approximated by:

$$p(\mathbf{u}) \approx \int I'(\mathbf{x}) \mathbf{dl}_{\mathbf{u}} \tag{4.4}$$

With $\mathbf{l}_{\mathbf{u}}(\mathbf{x}) = \mathbf{P} \cdot \mathbf{x}$ the ray connecting the point $\mathbf{u} \in \mathbb{R}^2$ on the detector plane to the emission source.

Eqn. (4.4) shows that $p(\mathbf{u})$ is invariant to shifts of the distribution of $I(\mathbf{x})$ along the path of the ray, as long as the integral of $I(\mathbf{x})$ is preserved. This means that displacements collinear to the projection rays cannot be directly observed in the projection image, since such displacement do not incur changes in $\int I'(\mathbf{x}) \mathbf{dl}_{\mathbf{u}}$, leaving the intensity of the projected image unchanged. A direct consequence of this result is that, in fluoroscopic images or DRRs, it is impossible to recover displacements along the direction of projection.

4.2.3 Network architecture

Our method relies on a fully convolutional architecture derived from ResNet ((K. He *et al.*, 2016)). The key characteristic of our network is the direct translation from a dense 2D input to a dense 3D output (see Fig. 4.2). The encoder part of our network produces 2D feature maps that are transformed by a backprojection module into 3D feature maps. The backprojection module samples the 2D feature maps using a regular grid in voxel space to obtain 3D feature maps, as described in

Sec. 4.2.3.1. This module uses the projection matrix $\mathbf{P}$ to relate spatial locations in the 2D feature maps to locations in the 3D feature maps. The decoder part of our network then transforms the 3D feature maps into a displacement field $\hat{\boldsymbol{\varphi}}$.

In the encoder, the first layer transforms the grayscale 256×256 input image into feature maps of shape $96 \times 128 \times 128$. Then, five ResNet blocks composed of two convolutional layers, each followed by a BatchNorm and a PReLU activation, process the data, with each block dividing the spatial resolution and multiplying the number of features by two. Finally, the last convolutional layer is applied without changing the number of features or resolution, to obtain 2D feature maps of shape $1536 \times 8 \times 8$. After the encoder, the backprojection module transforms the features from 2D projective space to 3D anatomical space, as described in Sec. 4.2.3.1, to obtain 3D features of shape $768 \times 4 \times 2 \times 4$, conserving the total number of features in the process. In the decoder, 10 deconvolutional layers are employed to decode these 3D features into a 3D displacement field of shape $3 \times 128 \times 64 \times 128$. These layers are not set up in ResNet blocks, with the number of features and spatial resolution being halved and doubled every two layers, respectively, and the last layer transforming the 24-dimensional feature map into a 3-dimensional displacement field.

During training, a reprojection loss $\mathcal{L}_{\boldsymbol{\varphi}^{2D}}$ and a 2D dice loss $\mathcal{L}_{s^{2D}}$ are used to optimize the network parameters in a supervised manner, as described in Sec. 4.2.3.2.

The network is implemented in PyTorch 2.4.1, with a memory footprint of ~ 335 MB. A forward pass is computed in 12 ms on an Nvidia RTX 4090 GPU. The Adam algorithm ((Kingma *et al.*, 2014)) was used to optimize the network weights. The network was trained for 5 epochs, using the One Cycle learning rate schedule ((Smith *et al.*, 2019)) to vary the learning rate η, with $\eta_{max} = 5.10^{-3}$. Training the network for more epochs (and a proportionally larger scheduler step size) did not lead to increased performance.

4.2.3.1 Backprojection module

Inspired by the LiftReg network ((L. Tian *et al.*, 2022)), we designed a backprojection module to handle 2D-3D spatial correspondence. An important difference between the LiftReg network and our approach is that we backproject the feature maps instead of backprojecting the input image. This largely reduces the number of parameters and inference time of the network because the encoder is composed of 2D convolutional layers instead of 3D convolutional layers.

Depending on the pose of the C-arm, the input projection image may be oriented arbitrarily with respect to the preoperative CT volume. The goal of this module is to ensure that the value of the displacement field at a voxel $\mathbf{x}$ will be determined by the value of the pixel at the coordinate $\mathbf{u} = \mathbf{P}\mathbf{x}$.

The backprojection module (in green in Fig. 4.2) contains no trainable param-

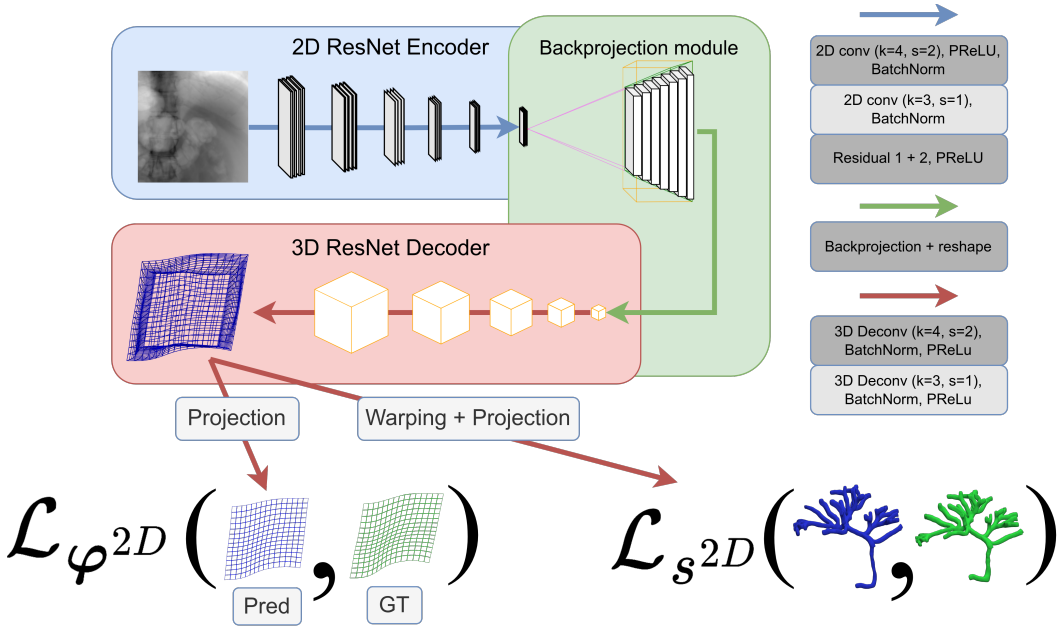


Figure 4.2: The *encoder* part of the network compresses the image into a low-resolution, high-dimensional feature space. The 2D features are backprojected to 3D using the projection matrix in the *backprojection module*. In the *decoder*, the network upscales the feature maps into a 3D displacement field. After the network, the loss is computed on the displacement of grid points projected in 2D space and on the warped vessels segmentation, projected in 2D space.

eters and relies on the projection matrix $\mathbf{P}$, obtained from the pose of the C-arm camera, to sample 3D features from 2D features via trilinear interpolation. The 2D feature map F is a tensor of shape (N, C, H, W), which can be treated as a batch of N 3D images with (C, H, W) dimensions. Since convolutions are approximately local operations, the (H, W) dimensions span the same spatial extent as the input 2D image. The channel dimension C is treated as a pseudo depth dimension, where each channel represents a different image plane. This is the first step in Fig. 4.3. It is important to note that these 3D features are not in the same space as the preoperative 3D anatomical image, because the input of the network is a perspective projection of the anatomy. Thus, these features must be mapped to the preoperative anatomical space by performing a backprojection operation. First, a regular grid of 3D points $\mathbf{G}$ is sampled in the bounding box of the region covered by the projection. This is the second step in Fig. 4.3. The grid $\mathbf{G}$, of shape $(3, C, H, W)$, can be converted to 2D coordinates using the projection matrix $\mathbf{P}$ by computing $\mathbf{G_u} = \mathbf{PG}$. To index the pseudo-depth dimension C of F, the first coordinate of $\mathbf{G_u}$ is set to an array of values varying from 0 to 1 along C.

Finally, F can be interpolated using $\mathbf{G_u}$ to transform the 2D projective features into 3D features F_{3D} aligned with the preoperative anatomical space. Elements of F_{3D} outside of F are simply set to 0 since they are not visible in the projection. This is the last step in Fig. 4.3. F_{3D} is further transformed by splitting the first dimension into a spatial and a channel dimension, to form a 5D feature map of shape (C_{3D}, D, H, W). After the backprojection, F_{3D} is input in the decoder (in red in Fig. 4.2) to predict a displacement field in the same space as the preoperative anatomy.

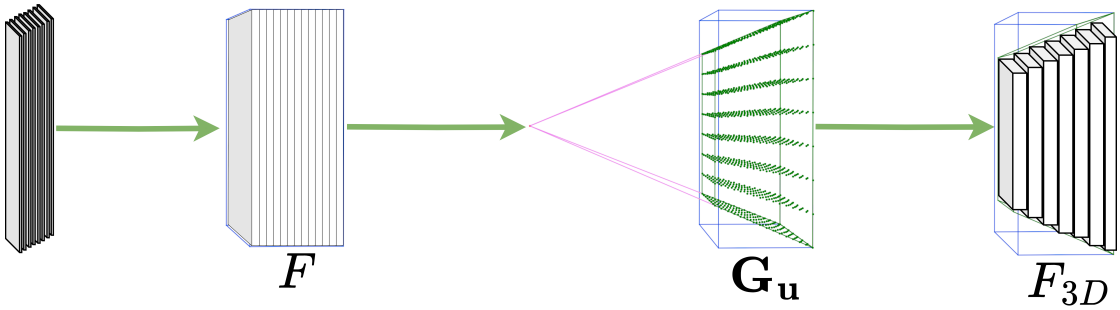


Figure 4.3: The 2D feature maps are reshaped into a volume in ray space. The volume is then sampled using ray space coordinates to obtain a volume in voxel space. Volume elements that are outside of the projection are set to zero.

4.2.3.2 Loss computation

Before computing the loss, we first mask the prediction and ground truth to exclude locations that are not visible in the image or outside the body. Because of the finite size of the detector, p is restricted to the domain Ω_p, defined by the size of the C-arm detector. This implies that the network can only predict $\hat{\boldsymbol{\varphi}}$ at the points $\mathbf{x}$ that have a projection in the 2D domain Ω_p: $\{\mathbf{x} \mid \exists\, t \in \mathbb{R}, \exists\, \mathbf{u} \in \Omega_p, t\mathbf{P}\mathbf{x} + \mathbf{x_0} = \mathbf{u}\}$ with $\mathbf{x_0}$ the origin of the C-arm camera matrix. Thus, we restrict the computation of $\mathcal{L}$ to the voxels that have a projection in Ω_p. Furthermore, the displacement at locations outside the body cannot (and should not) be recovered by the network due to the lack of contrast at those voxels, which typically contain air. Thus, we build a body mask using TotalSegmentator ((Wasserthal *et al.*, 2023)), and warp it with the ground truth displacement field to mask out locations that are not in the deformed body.

We formulate the loss function $\mathcal{L}$ as combination of two losses:

$$\mathcal{L} = \mathcal{L}_{\boldsymbol{\varphi}^{2D}}(\boldsymbol{\varphi}_i^{2D}, \hat{\boldsymbol{\varphi}}_i{}^{2D}) + \lambda \mathcal{L}_{s^{2D}}(\mathcal{P}_{2D}(I_s \circ \boldsymbol{\phi}_i), \mathcal{P}_{2D}(I_s \circ \hat{\boldsymbol{\phi}}_i)) \tag{4.5}$$

$\boldsymbol{\varphi}_i^{2D}$ and $\hat{\boldsymbol{\varphi}}_i{}^{2D}$ are the ground truth and predicted displacements, respectively, projected in the image plane. $\mathcal{P}_{2D}(I_s \circ \boldsymbol{\phi}_i)$ and $\mathcal{P}_{2D}(I_s \circ \hat{\boldsymbol{\phi}}_i)$ are the ground truth

and predicted segmentations, projected in the image plane. $\mathcal{L}_{\varphi^{2D}}$ is the Mean Square Error (MSE) on the projected displacement, and $\mathcal{L}_{s^{2D}}$ is the Soft Dice loss ((Milletarì *et al.*, 2016)) on the projected segmentation.

To compute $\mathcal{L}_{\varphi^{2D}}$, a 3D grid of points defined on the field domain is first warped with the predicted and ground truth displacement fields. Then, the warped 3D points are projected using $\mathbf{P}$ to obtain warped 2D points. Finally, $\mathcal{L}_{\varphi^{2D}}$ is computed between the ground truth and predicted 2D points.

To compute $\mathcal{L}_{s^{2D}}$, the vessels segmentation is warped with the predicted and ground truth displacement fields. Then, the fully differentiable renderer DiffDRR ((Gopalakrishnan and Golland, 2022)) is used to obtain 2D projections of the warped segmentations. Finally, $\mathcal{L}_{s^{2D}}$ is computed between the ground truth and predicted 2D segmentations.

This novel combination of loss in 2D image space is better suited to the clinical objective of visualizing 3D preoperative data registered on 2D intraoperative images. In contrast, supervising the network in 3D anatomical space does not take into account the fact that there is a loss of information in the 2D image formation process (see Sec. 4.2.2.3). Furthermore, since our objective is not 3D volume reconstruction but rather augmented fluoroscopy, we do not supervise on 3D image intensity values. The performance impact of this loss over alternative formulations is studied in Sec. 4.5.3.7.

4.2.4 Data augmentation

To improve the network robustness to changes in input image appearance unrelated to deformations, such as the 'sim-to-real' domain gap, we used a data augmentation method described in ((Grimm *et al.*, 2021), post-processing section). This data transformation is a composition of perturbations, each successively applied at random to the input with a probability of 50%. Perturbations include noise addition, smoothing, and contrast change. In ((Grimm *et al.*, 2021)), authors demonstrate that this approach enables a 2D-3D rigid registration network trained on DRRs to perform well on real fluoroscopic images. In this work, we find that this data augmentation scheme improves the generalization capabilities of the network (see Sec. 4.5.3.9).

4.2.5 Data processing

To improve the training of our deep neural network and generate appropriate ground truth data for the loss function detailed in Sec. 4.2.3.2, several processing steps are required. These steps focus on properly integrating 2D-3D geometrical information into the training process. They are performed during training initialization (Sec. 4.2.5.1), before the network prediction (Sec. 4.2.5.2), and prior to loss

computation (Sec. 4.2.5.3).

4.2.5.1 Training initialization

First, the DRR projection parameters, CT volume geometry, and segmentation volume used to compute $\mathcal{L}_{s^{2D}}$ are loaded. The grid of coordinates used to compute $\mathcal{L}_{\varphi^{2D}}$ is then created in the field domain. At this stage, the input data transformation is also initialized, which includes optional operations such as cropping, rotation, padding, flipping, standardization or normalization, resizing, and data augmentation (Sec. 4.2.4). By default, this transformation rotates the input 90° counterclockwise, such that DRRs correspond to the projection matrix, applies standardization (subtracting the dataset mean and dividing by its standard deviation), resize the input to $(256, 256)$, and applies data augmentation. Experiments revealed that a $(256, 256)$ resolution provided better performance compared to other resolutions, and standardization, a common machine learning practice, also slightly improved results. When resizing, the projection matrix is adjusted by scaling the focal length and optical center to match the new image dimensions. At this point, the 3D coordinates $\mathbf{G}$ and 2D coordinates $\mathbf{G_u}$, used in the backprojection module, are computed.

We also generate an input mask to exclude parts of the image that show volume regions outside the field domain. This is achieved by tracing rays between a grid of 2D image points and the center of projection using the projection matrix and calculating the rays' intersections with the planes delimiting the field domain. Points corresponding to rays that do not pass through the field domain or cross 'forbidden planes' are added to the mask, which ensures that discontinuities in the displacement field remain invisible in the image (as illustrated in Fig. 4.4). The input mask can be combined with additional masks, such as to replicate the X-ray beam collimation of testing data.

From the 2D input mask, a corresponding 3D mask is generated for use in the backprojection module and before loss computation. This 3D mask is created by sampling the 2D mask at each coordinate $\mathbf{G_u}$, to verify whether the projection of each $\mathbf{G}$ point falls within the mask. Finally, the segmentation volume used in $\mathcal{L}_{s^{2D}}$ is loaded.

4.2.5.2 Data pre-processing

Before each prediction, the input data transformation and input mask are applied to the input DRR. The coordinates $\mathbf{G}$ are then warped using the ground truth displacement field to sample the body segmentation at deformed coordinates. As mentioned in Sec. 4.2.3.2, the deformed body segmentation is added to the 3D mask to remove voxels outside the body from the loss computation.

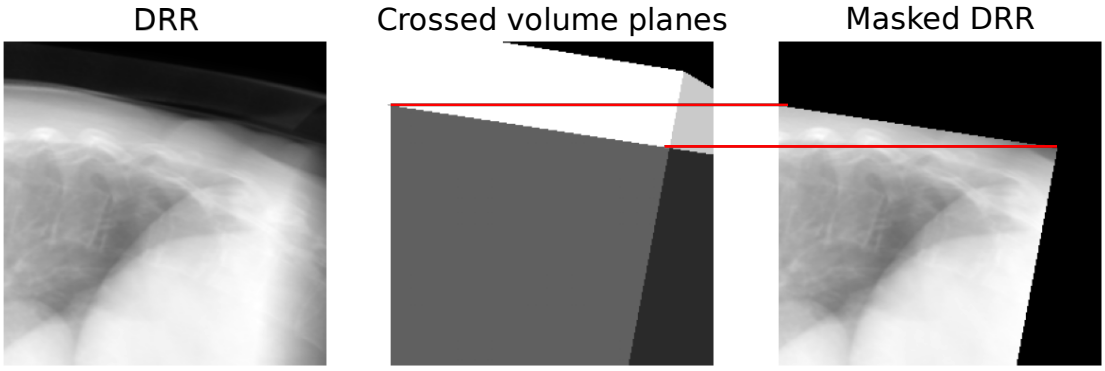


Figure 4.4: From left to right, a DRR generated with a randomized pose, a visualization of planes crossed by projections rays, the DRR with the input mask applied. To visualize the crossed planes, each plane was assigned an integer ID, such that each sum of two planes had a unique value, and the sum of crossed planes was plotted for each ray. The red lines were added to make the interpretation of the figure easier and underline that regions where the rays cross forbidden planes are removed. In the DRR on the left, we can see that these forbidden regions correspond to regions where the 'outside' of the CT volume is partially visible.

4.2.5.3 Data post-processing

Before computing the loss, the 3D mask is applied to both the predicted and ground truth displacement fields to ensure that subsequent computations are restricted to regions in the mask. Then, the ground truth and predicted warped 2D segmentation $\mathcal{P}_{2D}(I_s \circ \boldsymbol{\phi_i})$ and $\mathcal{P}_{2D}(I_s \circ \hat{\boldsymbol{\phi}}_{\boldsymbol{i}})$ are rendered by tracing rays from the center of projection to the projection plane through the segmentation volume. In the rendering process, rays are traced from the center of projection to the projection plane through the segmentation volume. For each ray, $N_s = 64$ points are uniformly sampled within the segmentation bounding box, avoiding regions outside the segmentation to conserve memory and increase resolution. These points are then warped by the ground truth or predicted displacement fields, and the segmentation is sampled at the warped positions. The final rendered segmentation, s^{2D} (or $\hat{s}^{2D}$) is computed by summing the sampled values along each ray. This process is similar to the DiffDRR (Gopalakrishnan and Golland, 2022) rendering procedure, with the addition of an intermediate warping step.

Finally, the input mask is applied to both s^{2D} and $\hat{s}^{2D}$, and $\boldsymbol{\phi_i}$ and $\hat{\boldsymbol{\phi}}_{\boldsymbol{i}}$ are converted from voxels to millimeter for the loss computation.

4.3 Domain-agnostic data generation enables robust registration

4.3.1 Introduction

Deformable registration between 2D and 3D medical images remains a critical challenge in image-guided interventions. While significant advances have been made in radiotherapy applications (Shieh *et al.*, 2017; Zhang, X. Huang, and Wang, 2020; Hirai *et al.*, 2019; M. D. Foote *et al.*, 2019), where statistical models derived from 4D CT scans effectively handle respiratory motion (see Sections 3.1.2 and 3.3.1), there is a pressing need to address more complex, intervention-specific deformations.

Surgical tumor removal, a primary cancer treatment modality, presents unique challenges for deformable registration. Unlike respiratory motion in radiotherapy, surgical deformations are largely unpredictable, arising from direct tissue manipulation by clinicians. These arbitrary deformations cannot be captured by conventional statistical models based on preoperative imaging. Currently, no 2D-3D deformable registration method adequately addresses the real-time tracking of tumor position under such conditions. Thus, the objective of this study is to propose an approach that overcomes these limitations by estimating diffeomorphic displacement fields in real-time without relying on statistical deformation models.

Our approach uses fluoroscopy as the intraoperative imaging modality to address scenarios requiring real-time deformation recovery. To reduce the need for preoperative 4D CTs of the patient, and to make the method as generic as possible, we propose a domain randomization solution to generate displacement fields using only a 3D preoperative CT. Our data generation is based on the large deformation diffeomorphic metric mapping (LDDMM) framework (Sec. 4.3.2.2) to enforce smooth and invertible displacement fields. As in previous works, we generate DRRs to simulate intraoperative fluoroscopic images (Sec. 4.3.2.2). These images serve as input to a ResNet-based fully convolutional network that learns the mapping between DRRs and 3D displacement fields (Sec. 4.3.2.1). The computed deformation field enables real-time tracking of internal structures, such as tumors, by warping preoperative CT data or derived 3D meshes (Sec. 4.3.3).

4.3.2 Preliminary approach to domain-agnostic registration

This study employs the method presented in Sec. 4.2 with some modifications, summarized below. The data generation and loss computation processes are unique to this work.

4.3.2.1 Network architecture

In this work, the network architecture presents the following differences with the architecture detailed in Sec. 4.2.3:

- The network input is a DRR of the deformed anatomy subtracted with the DRR of the preoperative anatomy.
- The network doesn't yet employ backprojection but rather a reshape operation (denoted as the 'NoBackproj' variant in Sec. 4.5.3.6) to transform 2D features to 3D features, which is less accurate.
- 'Adversarial Noise Layers' (ANL) (You *et al.*, 2019) are employed in the network after the activation functions, with the goal to regularize the latent space. However, as we later found that these layers did not bring performance improvements, we did not use ANL in further experiments.
- In this work, an MSE between the ground truth displacement field and the predicted displacement is used to optimize the network parameters, with the displacement in the direction perpendicular to the projection set to 0.
- No learning rate scheduler is used.
- The number of filters in the first convolutional layer is 64 instead of 96.
- 6 encoder and decoder layers are used instead of 10
- The shape of the predicted and ground truth displacement fields is $(64, 32, 64)$ instead of $(128, 64, 128)$.
- The domain randomization data augmentation described in Sec. 4.2.4 was not used.

4.3.2.2 Data generation

To train our network, we generate a synthetic dataset composed of DRRs paired with 3D displacement fields, which are used to deform the preoperative CT. From the deformed CT Scans, we generate DRRs using the DeepDRR algorithm (Unberath *et al.*, 2018). We detail below these different steps and motivate our choices.

Deformation generation: The general framework behind generating and applying deformations is detailed in the first paragraph of Sec. 4.2.2.1. The exact process used to generate the synthetic training dataset in this work is different from the process in Sec. 4.2.2.1, and is detailed below.

Domain randomization: First, we recall that, as in Sec. 4.2.2.1, the displacement field $\boldsymbol{\varphi}(\mathbf{x})$ is computed on a grid of points $\mathbf{x}$ in the field domain via Eqn. 4.1. However, instead of generating displacement fields in one step, $\boldsymbol{\varphi}(\mathbf{x})$ is computed iteratively. At each step t, random $\boldsymbol{\sigma_k}(t)$ and $\boldsymbol{\alpha_k}(t)$ are sampled around initial (randomly set) mean values and the corresponding displacement field $\mathbf{v}(t)$ is computed from these parameters. Then, the deformed points $\mathbf{x}(t+1) = \mathbf{x}(t) + \mathbf{v}(t)$ are computed and the process is repeated until reaching a set number of steps.

With this process, the total displacement field is given by $\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{x}(t_{max}) - \mathbf{x}(t_0)$. Computing $\boldsymbol{\varphi}(\mathbf{x})$ in this way provides takes advantage of theoretical guarantees demonstrated by the LDDMM framework (Trouve *et al.*, 2005). It is also possible to numerically verify that the displacement field computed at each step is a diffeomorphism by checking that $\|v_t(\mathbf{x}_t)\|_{W^{1,\text{inf}}} < 1$ where $\|v_t\|_{W^{1,\text{inf},(\mathbb{R}^N,\mathbb{R}^N)}} = sup_{\mathbf{x}\in\mathbb{R}^N}(|v_t(\mathbf{x}_t)|_{\mathbb{R}^N} + |\nabla v_t(\mathbf{x}_t)|_{\mathbb{R}^N\times\mathbb{R}^N})$ which ensures that the final displacement $\boldsymbol{\varphi}(\mathbf{x})$ is a diffeomorphism as well (Allaire, 2006).

However, in practice, using this process over the process detailed in Sec. 4.2.2.1 didn't significantly improve the quality of the generated displacement fields. Thus, due to the cost associated with computing $\boldsymbol{\varphi}(\mathbf{x})$ iteratively as well as the resulting lack of closed form solution for $\boldsymbol{\varphi}(\mathbf{x})$ with this approach, the process detailed in Sec. 4.2.2.1 was used in later works. Another notable difference with the deformation generation process in this work is that no global scaling is applied to $\boldsymbol{\varphi}(\mathbf{x})$ with the consequence that small global deformations are not appropriately represented in the dataset.

PCA-based deformations: In clinical scenarios when a 4D-CT is available preoperatively, it can be beneficial to leverage an *a priori* knowledge about the type of deformations that the organ undergoes. Using a registration method based on the LDDMM framework, such as SyN (B. B. Avants *et al.*, 2008) implemented in ANTs (Brian B. Avants *et al.*, 2011), we can compute a set of DVFs from the 4D-CT. Following the idea proposed in Chou *et al.* (C. R. Chou *et al.*, 2012) and Foote *et al.* (M. D. Foote *et al.*, 2019), we compute a PCA over the time series of diffeomorphic deformations obtained from the 4D-CT. A new set of deformation fields is then generated by sampling the PCA deformation subspace. We keep the 3 first PCA components (which explain 90% of the variance). The associated weights are randomly sampled between ±150%, ±130% and ±110% of their respective maximum value to generate the DVF dataset.

4.3.3 Results & discussion

We present here an evaluation of our results on a series of lung CTs. Using the 4D-CT we can compare the 2 variants of our method, i.e. the PCA-based and the

Parameter considered	Validation loss value
$(H, W) = (128, 128)$; $(H_{out}, D_{out}, W_{out}) = (32, 16, 32)$	49.22
$(H, W) = (256, 256)$; $(H_{out}, D_{out}, W_{out}) = (64, 32, 64)$	38.41
$N_{conv} = 6$	38.41
$N_{conv} = 10$	53.69

Table 4.1: Ablation study resulting in the following set of parameters offering the best compromise between computation time, memory usage and accuracy: $N_{conv} = N_{deconv} = 6$, $C_0 = 64$, $H = W = 256$, $H_{out} = W_{out} = 64$, $D_{out} = 32$.

Domain Randomization based data generation techniques, on increasingly large deformations.

Experimental setup: The validation dataset (Hugo *et al.*, 2017) includes 10 CT volumes ($512 \times 512 \times 142$ voxels) with a voxel size of $0.98 \times 0.98 \times 3\,\text{mm}^3$. In each volume we selected a region of interest (ROI) large enough to contain the lungs, with additional margins to accommodate anatomical displacements. The dimensions of the ROI are $243 \times 164 \times 79$ voxels. The DVFs in the Domain Randomization dataset were computed at a resolution of $64 \times 32 \times 64$. The basis displacements were first computed at full resolution and then downsampled to a $64 \times 32 \times 64$ size before applying the PCA transform. The corresponding DRR were generated with an initial resolution of 960×960 pixels and were subsequently downsampled to a 256×256 size. The DRR corresponding to the undeformed CT volume was subtracted to all the DRRs in the dataset to better correlate the information content in the image with the information from the DVF. A total of 20,000 samples were used to train our network for 45 epochs, with a learning rate initialized at 5.10^{-5} and multiplied by 0.9 each time the validation loss would plateau for 5 epochs. The training converged in 37 hours on an Nvidia GeForce GTX 1080 Ti. Table 4.1 presents the optimal hyperparameters found for our network.

Results: Our results are summarized in Table 4.2. We can quickly see that using the added information available in the 4D-CT provides better registration results on both the full image and TRE. This is similar to what can be observed in (M. D. Foote *et al.*, 2019), and it is no surprise, as the training data is very close to the problem characteristics. Yet, in this context, our approach leads to better results than the state of the art. In Foote *et al.* (M. D. Foote *et al.*, 2019), the max registration error on the time series is 9.55 mm, whereas our method achieves a max error of 2.22 mm.

Phase #	*PCA-based variant* mean (max) error	*DR-based variant* mean (max) error	*PCA-based variant* mean TRE	*DR-based variant* mean TRE
0	0.16 (0.95)	0.57 (2.04)	0.0	0.8
1	0.16 (1.03)	0.94 (3.54)	1.52	1.85
2	0.28 (1.69)	1.39 (6.14)	1.24	3.02
3	0.28 (2.22)	1.64 (7.26)	2.11	2.97
4	0.25 (1.4)	1.92 (9.05)	0.86	2.96
5	0.28 (1.68)	2.08 (9.09)	1.65	2.76
6	0.21 (1.13)	2.12 (10.78)	1.38	3.16
7	0.3 (2.2)	1.76 (8.27)	1.12	1.97
8	0.25 (1.83)	1.25 (5.06)	0.72	1.84
9	0.21 (1.51)	0.76 (2.74)	1.31	1.71

Table 4.2: Summary of our results. The table on the left presents the registration errors on the entire volume, for the problem optimized data set generation (column 2) and the generic training data set based on domain randomization (column 3). The table on the right presents mean target registration errors (TRE) for the same variants of our method. Note that each row corresponds to a phase of the respiratory cycle, and therefore rows number #4 and #5 have the largest deformation.

However, since a 4D-CT is needed to obtain such results, this method may only be applied to a reduced set of clinical scenarios. For all other scenarios where only a preoperative CT is available, table 4.2 shows that our Domain Randomization method performs very well (average registration error below 2.5 mm and TREs below 3.2 mm) while trained on completely generic displacement fields.

Discussion: The objective of this work was to propose an accurate real-time 2D-3D registration method for fluoroscopy-guided interventions without fiducial markers. We show that this ill-posed problem can be solved via deep learning when associated with an efficient data generation pipeline. Our method leverages Domain Randomization combined with a diffeomorphic displacement field generation, and only requires routinely acquired images as opposed to other methods in the state-of-the-art, thus improving its applicability to various clinical settings. Our results show that the proposed method can estimate a 3D displacement field, even for structures deep into the tissues, with an average accuracy of 1.44 mm. We also show that our framework can be extended leverage 4D-CT preoperative data. Using the additional displacement information contained in such time series, we

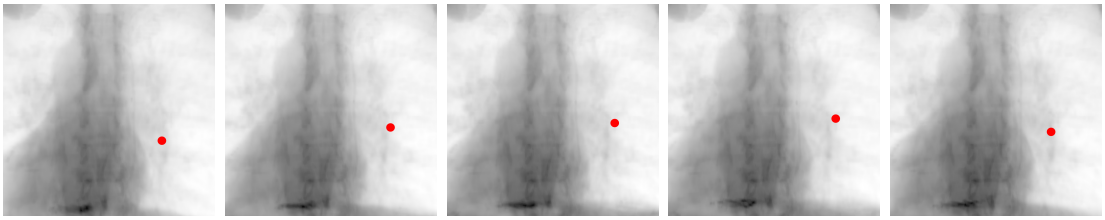


Figure 4.5: Illustration of a tumor tracking over a respiratory cycle. Images, from left to right, correspond to phases 0, 2, 4, 6, and 8 of the time series described previously. The tumor location is accurately updated at a frequency of 20Hz.

obtain an even higher accuracy of 0.24 mm. This level of accuracy is obtained at an update rate of about 20 Hz, sufficient for all clinical applications. An illustration of tumor tracking during respiratory motion is illustrated in figure 4.5.

4.4 Vessel deformation prediction without contrast agents

4.4.1 Introduction

The evolution of image-guided procedures has transformed medical imaging from a purely diagnostic tool into an essential component of therapeutic interventions. This transformation has given rise to specialized fields such as interventional radiology, therapeutic endoscopy, and minimally invasive image-guided surgery, where improved outcomes have driven widespread adoption (Epstein *et al.*, 2013). X-ray-based imaging, particularly fluoroscopy, serves as the cornerstone of these procedures, offering real-time visualization capabilities crucial for precise navigation. However, the current reliance on contrast agents for vessel visualization presents significant clinical challenges. The nephrotoxicity of intravascular contrast media limits the injectable volume (Mamoulakis *et al.*, 2017), while the transient nature of contrast enhancement creates procedural inefficiencies and interruptions in real-time guidance. These limitations highlight a pressing need for contrast-free vessel visualization methods that can maintain continuous, real-time tracking of vascular structures during interventions.

This study addresses this clinical need and evaluates the use of our proposed deep learning approach for predicting vessel deformation in fluoroscopic imaging without contrast enhancement. The objective of this work is to enable continuous, real-time vessel visualization during fluoroscopy-guided procedures, potentially improving both safety and efficiency while reducing the procedural dependence on contrast agents. We demonstrate our method's capability to predict respiratory-induced hepatic vein deformations from synthetic fluoroscopic images, with detailed results presented in Sec. 4.4.2.

The data generation process employed in this work is identical to the data generation process described in the previous work (Sec. 4.3.2.2). The network architecture presents the following differences with the architecture detailed in Sec. 4.2.3:

- The network input is a DRR of the deformed anatomy subtracted with the DRR of the preoperative anatomy. In future studies, we remove this subtraction operation since it did not lead to improved performances.
- The network doesn't yet employ backprojection but rather a reshape operation (denoted as the 'NoBackproj' variant in Sec. 4.5.3.6) to transform 2D features to 3D features, which is less accurate.
- In this work, an MSE between the ground truth displacement field and the predicted displacement is used to optimize the network parameters, with the

displacement in the direction perpendicular to the projection set to 0.

- No learning rate scheduler is used.
- The number of filters in the first convolutional layer is 64 instead of 96.
- 6 encoder and decoder layers are used instead of 10
- The shape of the predicted and ground truth displacement fields is $(64, 32, 64)$ instead of $(128, 64, 128)$.
- The domain randomization data augmentation described in Sec. 4.2.4 was not used.

4.4.2 Results

A human liver CT obtained from a patient of the Paul Brousse hospital in Paris was used to generate a 10,000 sample dataset, split into 8,000 training samples and 2,000 validation samples. The maximum amplitude of deformation in the dataset was 22 mm and 40 mm in the LR direction and SI direction respectively.

The testing dataset was generated from the same 3D CT, this time using BSpline transforms tailored to mimic a breathing motion. Specifically, inhale and exhale phases and sliding motion of the organs against the bones were modeled. In this case, the maximum amplitude was 10 mm and 25 mm for the SI and LR directions. The dataset contains 5 inhale/exhale periods for a total of 50 samples.

The accuracy of the network was measured via the reprojection distance (RPD) metric. Hepatic veins were deformed using the ground truth and the predicted displacement fields. The deformed mesh points were projected onto the image plane and the 2D RPD error was measured.

The mean RPD error on the testing dataset was 2.7 ± 1.9 mm while the mean RPD displacement was 7.7 ± 3.9 mm. Figure 4.8 shows the distribution of the error on the hepatic veins. Figure 4.7 shows an example of image augmentation by our method. A full video is available at `https://mimesis.inria.fr/project/augmented-fluoroscopy/`.

4.4.3 Discussion

Even though the testing data were generated differently from the training data, the prediction of the network still reduced the error from 7.7 to 2.7 mm. These results validate our Domain Randomization approach, as the network learns to map a deformed fluoroscopy to a displacement field, and generalizes well on the testing data. This constitutes an advantage over other methods that might use

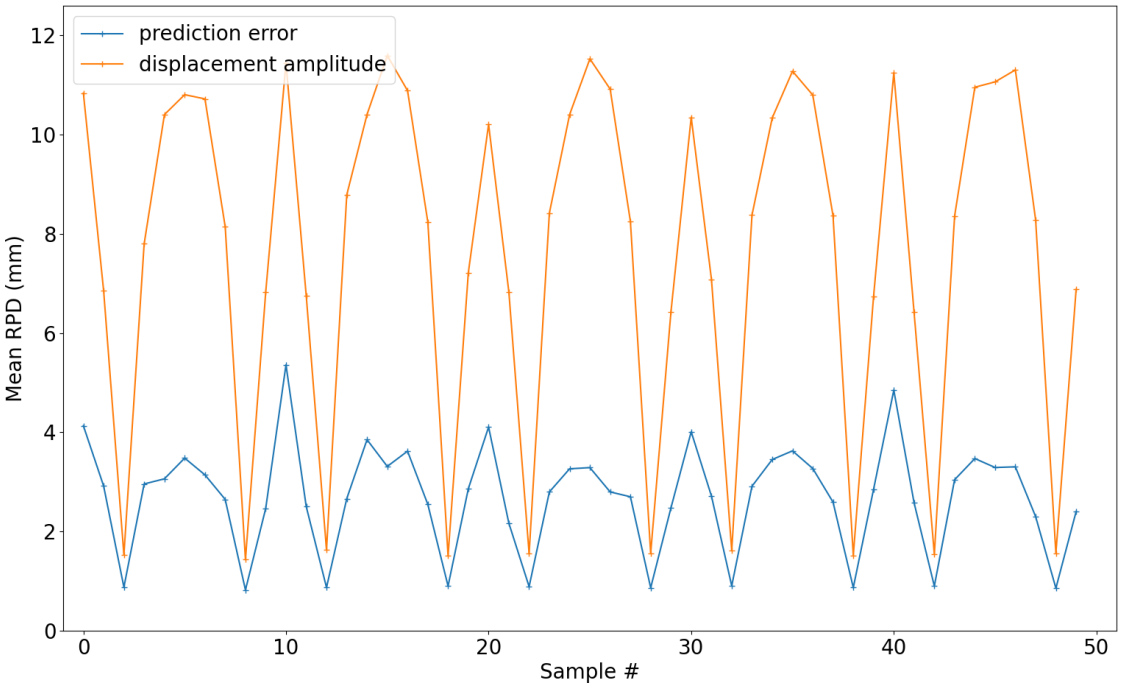


Figure 4.6: The average RPD error of our method on the testing data against the average RPD displacement.

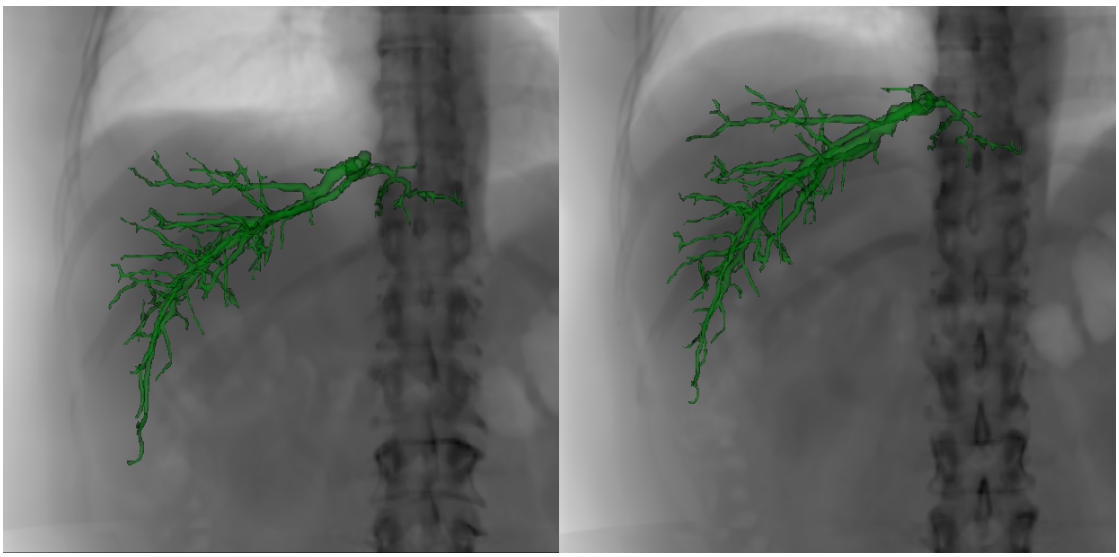


Figure 4.7: Augmented DRR at full inspiration (left), and full expiration (right), with the predicted hepatic veins position.

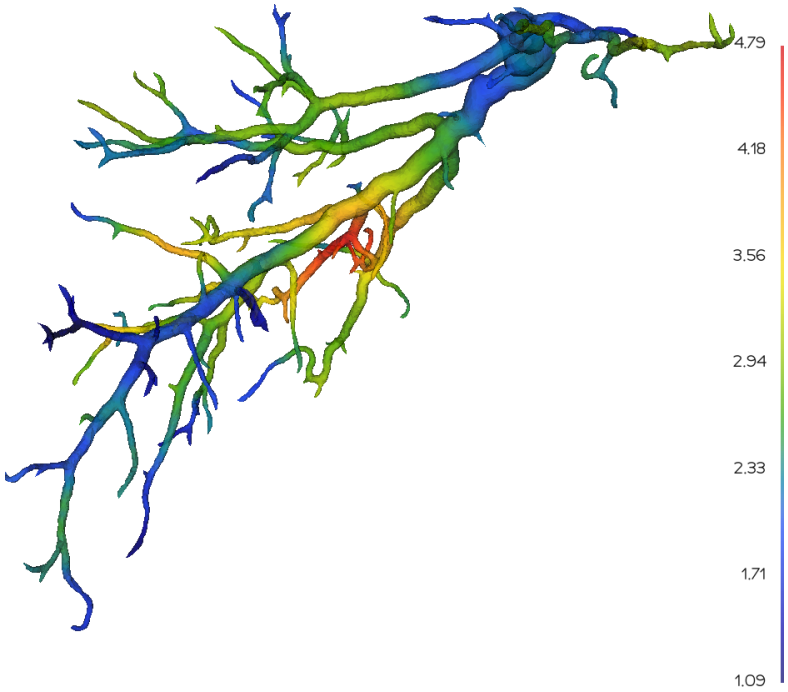


Figure 4.8: The average RPD error in mm on the hepatic veins position on the testing dataset.

a patient-specific motion prior obtained from a 4D-CT to train a neural network to predict deformations from a fluoroscopic image. A limitation arising from 2D fluoroscopy is that the displacement perpendicular to the image plane is not visible and thus cannot be predicted, but this is mitigated by the fact that the resulting error is also not visible on the augmented image.

4.5 2D-3D registration of intervention-related deformations

4.5.1 Introduction

The challenge of tracking anatomical structures during interventional procedures can be reduced to a 2D-3D registration problem, where real-time fluoroscopic images (2D) must be accurately mapped to preoperative CT volumes (3D) despite complex, intervention-induced tissue deformations. This registration problem is especially critical in liver interventions, where accurate tracking of vascular structures and tumors through tissue deformation directly impacts procedural safety and efficacy.

The liver, a complex solid organ, houses three distinct vascular systems and the biliary tree. It is unfortunately targeted by a wide variety of diseases, including tumors, both primary (affecting the liver itself) and secondary (resulting from diseases in other organs). The evolution of image-guided techniques has significantly improved diagnostic, theragnostic, and therapeutic procedures ((Epstein *et al.*, 2013)). These new techniques have also reduced invasiveness and enriched the information available for guiding the interventions.

Liver therapies guided by imaging techniques can be categorized into three primary groups of procedures: those targeting the liver parenchyma, those involving vascular structures, and those related to the biliary tree. The most performed procedures for each group are; for the **parenchyma**: liver or tumor biopsy, and tumoral ablation; for the **vessels**: portal or hepatic vein embolization, transjugular intrahepatic portosystemic shunt (TIPS), and transarterial chemoembolization (TACE); and for the **biliary system**: cholangiography, cholecystostomy, biliary drainage, endocanalicular biopsy, and stent placement.

X-ray-based imaging is one of the most frequently used imaging modalities for guiding the above-mentioned liver therapies. It includes CT, 2D fluoroscopy, and 3D fluoroscopy (i.e., cone-beam CT). However, low tissue contrast is a prominent limitation of X-ray-based imaging, requiring the injection of contrast agents (CA) into the bloodstream to enhance contrast between structures, enabling recognition, and thus, facilitating the procedures. During procedures, the liver is subject to deformations caused by cardiac and breathing motion, and the instrumentation performed by the clinician. This means that, in the absence of contrast injection, their position becomes indeterminate, with a lack of context, leading to risks of unintended damage. However, the volume of CA that can be injected during an intervention is limited due to its potential toxic effects when injected intravascular, notably nephrotoxicity ((Mamoulakis *et al.*, 2017)). Furthermore, the rapid dissipation of the contrast effect after injection makes the visualization of low-contrast

tissues only possible shortly after injection. Consequently, while fluoroscopy offers real-time capabilities, this transient contrast effect renders fluoroscopy most of the time an asynchronous image guidance method, available either before or after any gesture.

Therefore, the objective of this work is to predict the motion of intrahepatic structures (e.g., vessels, tumors, liver segments) in fluoroscopy-guided interventions without the need for CA injection. It is a challenging task due to the poor contrast between the structures and the surrounding tissue in fluoroscopic images. This task is further complicated in clinical settings where organs are deformed due to surgical manipulations or needle-based (percutaneous) interventions. Thus, we propose a *domain-agnostic* 2D-3D deformable registration method able to recover *arbitrary deformations* of the anatomy from a single fluoroscopic image. Our approach is based on the generation of a randomized, synthetic dataset to train a neural network that predicts deformations. Using a preoperative CT scan as a prior to generate a bespoke training dataset, our approach has the potential to provide real-time guidance in fluoroscopy-guided interventions while avoiding contrast agent injection, thus enhancing the safety and efficiency of procedures.

4.5.2 Previous works

Outside of the medical field, extracting 3D information from one or several 2D projective images is a widely encountered problem. Some works adopt a global approach, aiming to recover a 3D scene from a single 2D optical image (Yin *et al.*, 2022). Other works focus on reconstructing a 3D mesh of an object from one or more 2D projections (N. Wang *et al.*, 2018; Salvi *et al.*, 2020; L. Li *et al.*, 2021). In the medical field, 3D CT, along with MRI scans, are ubiquitous as they bring crucial information to clinicians. However, CT scans incur a non-negligible radiation dose for the patient and are not a real-time image modality. This is why a variety of works (Shen *et al.*, 2019; You Zhang, 2021; Lei, Z. Tian, T. Wang, Roper, *et al.*, 2021; C.-W. Chang *et al.*, 2022; Lei, Z. Tian, T. Wang, Axente, *et al.*, 2022) tackle the problem of 3D CT reconstruction from one or few 2D fluoroscopic images. Since reconstructing a 3D image from a few projections is an ill-posed problem, some of these works rely on a previously acquired CT image. These approaches are, however, not well adapted to our problem because they aim to reconstruct a full CT volume rather than update preoperative data for intra-operative visualization.

A similar problem is often tackled in radiotherapy, where the position of a tumor from the 3D preoperative CT scan must be updated to follow the breathing motion of the patient. This problem has been solved with success by tracking an implanted radio-opaque marker near the tumor in a bi-plane fluoroscopic image while the patient breathes ((Adler *et al.*, 1997; Seppenwoolde *et al.*, 2011)). How-

ever, marker implantation requires an additional procedure, potentially leading to complications. This is why more recent works (presented in Sec. 3.1) aim to directly predict the tumor location from two or more fluoroscopic images

Another, more general, approach is to predict a displacement field to update the position and shape of preoperatively segmented structures. Like our approach, such methods are referred to as 2D-3D deformable registration methods. A comprehensive state of the art of these methods is proposed in Sec. 3.3.

While previous works, particularly single-plane methods, have demonstrated the feasibility of 2D-3D fluoroscopy to CT real-time registration in radiotherapy, a critical gap remains in addressing arbitrary deformation recovery in other interventional procedures. These deformations result not only from respiratory motion but also from the mechanical influence of surgical instruments on tissue, leading to complex and unpredictable changes. Traditional approaches for tracking anatomical structures in fluoroscopy-guided interventions have relied on markers or contrast agents, posing limitations and potential risks. Meanwhile, all existing markerless 2D-3D tracking methods are based on pre-established motion models, suitable only for scenarios where the movement, such as periodic breathing, is known in advance. Thus, no solution exists for contrast-free, real-time tracking of anatomical structures in fluoroscopy-guided interventions.

This work directly addresses this unmet need by validating our domain-agnostic method for single-view 2D-3D deformable registration (section. 4.2) on an experimentally acquired porcine dataset with intervention-related deformations and two synthetic human datasets with intervention-related and breathing-induced deformations. Additionally, we performed a sensitivity analysis to evaluate the impact of the backprojection module and loss function on performance. Our results underscore the clinical relevance of our method, demonstrating real-time, contrast-free tracking of intrahepatic vessels, which could reduce the need for contrast agents in fluoroscopy-guided interventions. With this approach, we aim to contribute to the development of safer, more broadly applicable markerless and contrast-free tracking solutions for fluoroscopy-guided interventions.

4.5.3 Results

We present here several results to demonstrate the robustness, accuracy, and genericity of our method.

4.5.3.1 Datasets description

Our method was evaluated on three datasets:

1. A clinically acquired porcine dataset, with four pairs of CECTs, before and after intervention-related deformations.

2. A breathing motion dataset, composed of a baseline, clinically-acquired pre-operative CECT, and 50 deformed CTs replicating 5 breathing motion sequences.

3. A needle insertion dataset, composed of the same baseline CECT as the breathing motion dataset and 50 deformed CTs mimicking deformations induced by needle insertion.

These datasets are designed to evaluate the ability of our method to recover deformations that would occur during interventions, either due to breathing, or caused by the interaction between surgical tools and tissues. The testing data inputs are DRRs generated from the deformed CTs, with the deformed vessel segmentations serving as ground truth.

With the first dataset (the IHUdeLiver10 dataset, in Sec. 4.5.3.3), we evaluated our method on DRRs generated from CT scans acquired after interventions. In our experiments, we used four pairs of {*baseline; deformed*} CECTs experimentally acquired on four different porcine subjects, to illustrate the flexibility of our approach to recover various intervention-related deformations. Specifically, the four samples we used represent the following interventional scenarios: needle-tissue interactions, inter-fractional motion and laparoscopic surgery, respectively. As opposed to other publicly available datasets, this dataset allows us to evaluate the ability of our method to recover real, intervention-related deformations. A limitation of this dataset is that, due to contrast agent dissipation, the deformed intrahepatic vessels do not match the rest intrahepatic vessels in the number and length of vessels. We mitigate this issue by manually processing the segmentations to obtain vessel trees as similar as possible.

To completely eliminate this uncertainty, we further validate our method on a clinically acquired human CECT, with synthetic deformations (Sec. 4.5.3.4 and Sec. 4.5.3.5). We created two datasets to evaluate our method on both naturally occurring and intervention-related deformations. With the breathing motion dataset, our objective is to show that our method can compete with other works specifically designed for breathing motion recovery. The needle insertion dataset allows us to perform an accurate, quantitative assessment of our method on intervention-related deformations without experimental sources of errors. The process of generating these datasets is described below.

Breathing motion dataset: To create a semi-synthetic test dataset that represents breathing motion, we used a data generation process different from the training data generation process. First, the thorax bones and liver were segmented automatically using TotalSegmentator (Wasserthal *et al.*, 2023). A 3D grid of control points was also created, covering the image at a resolution of 20 mm in every

direction. A random displacement, biased to have a large vertical component, was generated uniformly for each control point. This displacement was then projected on the vectors tangential and orthogonal to the liver at each control point. Then, a distance map was built from the thorax bones segmentation and used to filter the displacement at control points. For each control point, if its distance to the bones was inferior to 10 mm, the tangential component was set to 0 and the orthogonal component was set to 10% of its value. This models both the sliding motion of organs against bones and outward motion during breathing. Finally, the full displacement field was scaled by a factor varying between 0 and 1 following a $\cos^4$ schedule to replicate a breathing pattern. To interpolate this displacement smoothly at every voxel, B-splines were used instead of linear interpolation to better handle the discontinuous nature of the displacement between control points. We generated five 10-phase synthetic 4D-CT with this process for a total of 50 samples. It is important to note that these deformations are generated in a completely different way than the training data, which is fully randomized, so our method is not biased to favor the reconstruction of these deformations.

Needle insertion dataset: Due to the action of surgical tools on the organ, intraoperative liver motion can be large, with displacements of up to 60 mm (Heizmann *et al.*, 2010). In this dataset, we simulated the insertion of a synthetic needle 2 mm in diameter 50 mm into the liver over the course of 50 time steps. The amplitude of displacement was 30 mm along the direction of needle insertion. This displacement was applied at control points inside the liver that were closer than 80 mm to the tip of the needle. With this approach, the deformation remained centered on the needle tip during insertion. A Deep Inspiration Breath Hold (DIBH) motion was also added, with the internal organs slowly moving upwards as the lungs slightly compress over the course of the breath hold.

4.5.3.2 Experimental setup

Sections 4.5.3.3, 4.5.3.4, and 4.5.3.5 present the results of our method on the IhuDeLiver10 dataset, the breathing motion dataset and the needle insertion dataset, respectively. A sensitivity analysis (section 4.5.3.6) was also performed to evaluate the impact of the proposed loss function and backprojection module on the IHUdeLiver10 dataset. Even though our method is agnostic to the choice of anatomical target structure, the portal vein tree was chosen as the target for both the porcine and human datasets, due to its clinical relevance, as introduced in Sec. 4.5.1. Here, the proposed application of our method is to replace contrast injection by superimposing the predicted vessel shape on fluoroscopic images. Thus, to evaluate the accuracy of our method, we measure the 2D Dice coefficient between the deformed and predicted vascular trees projected on the fluoroscopic image plane.

The parameters for the data generation were not tuned between the use cases and are common for all cases. The networks were trained from scratch for each case. The voxel resolution was different for each CT, but was always $< 1\,\text{mm}^3$. The DVFs were generated at a resolution of $128 \times 64 \times 128$ (in LR-AP-SI order) to reduce memory usage and processing time. To deform the baseline CT, the DVFs were upsampled linearly to the shape of the field domain before warping. The virtual C-arm was positioned automatically such that the liver was in the center of the projection, with a margin of $(100., 50., 100.)$ mm around the liver bounding box to ensure the liver is always visible in the projection. The DRRs were generated at a resolution of 512×512 pixels, with a pixel size of 0.67 mm, and downsampled to 256×256 before being input to the network. We did not observe performance gains when processing the input image at full resolution and downsampling reduced memory usage and computing time. The field domain was set to the bounding box of the projection in the volume. The number of random control points in the ROI was set to $N_{cp} = 30$. This number was chosen arbitrarily to provide enough variability in the generated displacement fields. We did not study the impact of this parameter on network performance. The norm of the generated displacement vectors was comprised between 0 and 100 mm, in order to cover sufficiently large deformations. For each baseline CT, $20,000$ samples were generated and randomly split into $18,000$ training samples and $2,000$ validation samples. Generating each training dataset took approximately 10 hours and 80 GB on a computer equipped with an Nvidia RTX 4090 GPU. Thus, due to time and space constraints, we did not generate a greater number of training samples.

4.5.3.3 Porcine dataset

The *IHUdeLiver10*[1] is a comprehensive collection of high-quality liver CT images, primarily focusing on organ deformations in 10 swine subjects undergoing various image-guided procedures. Specifically, four cases were selected from the dataset based on their procedures and the ensuing deformations. The deformation mechanism was needle-tissue interactions in subjects 1 and 2, inter-fractional motion in subject 3, and deformations before and after laparoscopic surgery in subject 4. Each case is composed of pairs of {*baseline; deformed*} multi-phase contrast-enhanced CT scans (MPCECT). Each MPCECT includes a non-contrasted phase and a portal-enhanced phase (after injection). Portal phases were processed by a liver surgeon with extensive knowledge of swine anatomy, including a detailed segmentation of the portal tree.

To remove contrast information, both the baseline and deformed scans were inpainted. As the contrast injections of baseline and deformed CTs were performed

[1]IHUdeLiver10 will be released at `https://doi.org/10.57745/EUBXGH`

Subject id	2D Dice	
	Registered	Baseline
1	0.651	0.416
2	0.694	0.558
3	0.655	0.278
4	0.593	0.476

Table 4.3: Accuracy of our method on the vessel tree registration for each subject in the porcine dataset.

at different times, the number and size of branches do not match. To alleviate this issue, we used the vascular modeling toolkit ((Izzo *et al.*, 2018)) plugin in 3D Slicer ((Fedorov *et al.*, 2012)), to compute the centerlines of each portal tree. We then voxelized the centerlines using a constant diameter of $\sim 5\,\text{mm}$. Finally, the same surgeon removed vessels that were absent from either the baseline or the deformed scan, ensuring that the Dice coefficient was computed between comparable structures.

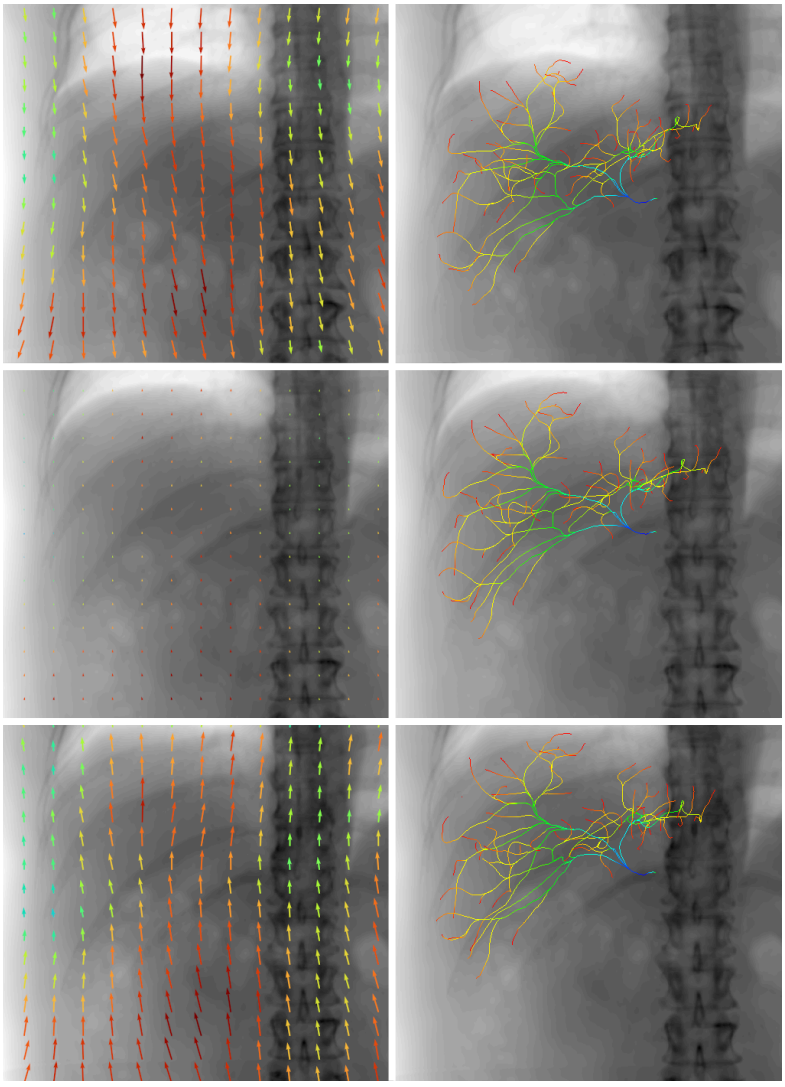
We evaluated the network's performance for each case by separately generating a training dataset and evaluating its performance on the deformed CT. We used the optimal parameters detailed in Sec. 4.5.3.6 for each case. We report the 2D Dice coefficient before and after registration in Table 4.3. On this dataset, we obtain a mean Dice of 0.649 ± 0.036, from 0.432 ± 0.102 before registration.

4.5.3.4 Breathing motion dataset

During an intervention, the deformation of the anatomy is either introduced by the clinician or by physiology, such as breathing. In the above experiment, we evaluated the ability of our method to recover deformations caused by the clinician's action on the anatomy. In this experiment, we evaluate the ability of our method to recover another important type of deformation occurring during interventions, respiratory motion. Fig. 4.9 shows an example of three phases in the testing dataset (left), overlaid with the vessel centerlines deformed by the network prediction (right).

We compare our method with the IGCN+ method (Nakao *et al.*, 2022), developed for 2D-3D breathing motion prediction. This approach uses a U-net CNN to extract the 3D motion of an organ as a 2D, 3-component deformation field. This 2D motion is then projected onto the organ mesh, and a Graph Convolutional Network is used to compute the motion on the occluded parts of the organ. In (Nakao *et al.*, 2022), the networks are trained and tested on a private 4DCT dataset composed of several cases. Here, we train IGCN+ following the imple-

mentation described in (Nakao *et al.*, 2022)[2], on a dataset generated from the first 4DCT in the breathing motion dataset. Specifically, we used a PCA on the first two motion components of the dataset as in (Nakao *et al.*, 2022) to generate the training dataset composed of ground truth 2D 3 components deformation fields and ground truth deformed organ meshes. In this case, the 'organ' on which the IGCN+ method was trained and evaluated is the portal vein tree.



[2]Original implementation available at `https://github.com/meguminakao/IGCN`

Figure 4.9: From top to bottom, DRR images of samples 0, 2, and 5 of the breathing motion dataset overlaid with the displacement field (left) and with the predicted vessel centerlines (right). The length and color of the arrows represent the magnitude of the displacement field. The vessel centerlines are colored for visualization purposes.

The performance of both methods was evaluated by measuring the 2D Dice coefficient between the ground truth and predicted portal veins segmentations. Fig 4.10 shows the performance of our method versus the IGCN+ method on the breathing motion dataset. On this dataset, our method obtained a mean Dice after registration of 0.863 ± 0.047 and the IGCN+ method obtained a mean Dice after registration of 0.879 ± 0.044, from 0.651 ± 0.074 before registration.

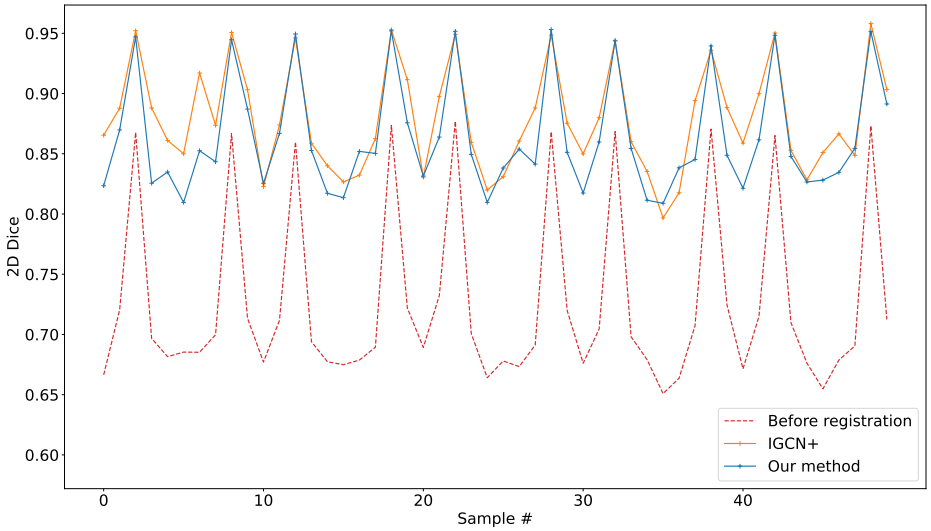


Figure 4.10: 2D Dice of our method and the IGCN+ method on the test set for each sample in the dataset. While the IGCN+ method obtains a slightly better performance on this problem thanks to the high similarity between its training and testing data, our method achieves an excellent result despite the lack of similarity between the training data and the testing data.

4.5.3.5 Needle insertion dataset

On this dataset, we evaluate the ability of our method to recover surgically induced deformations. In this dataset, the amplitude of displacement ranged from 0 to ≈ 30 mm over the course of the needle insertion. We used the same weights as in the

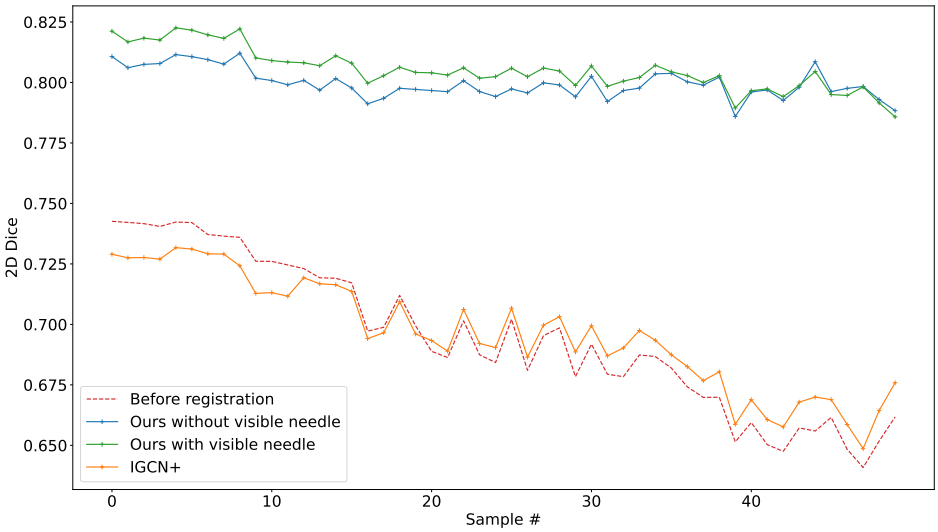


Figure 4.11: 2D Dice for each sample in the test set. Our method (blue) obtains similar performances on the needle insertion dataset and breathing motion dataset. In contrast, the IGCN+ method (orange) fails on this task due to being trained solely on breathing motion, performing only slightly better than no registration (red). Our method is also robust to the presence of surgical tools, such as needles, in the image when trained accordingly. The performance of our method is similar when trained and tested with needles in the image (green) and without (blue).

above experiment for both our network and the IGCN+ network. Fig 4.11 shows the performance of our method and the IGCN+ method on the needle insertion dataset. On this dataset, our method demonstrated very similar performances as in the breathing motion dataset, with a mean Dice after registration of 0.800 ± 0.006, while the IGCN+ performances were largely degraded, with a mean Dice after registration of 0.696 ± 0.023, nearly the same as before registration (0.688 ± 0.031).

We also evaluated whether our model could be trained to be robust to the presence of a needle in the image, as it occurs in percutaneous interventions. To train the network, we randomly overlaid needles of varying diameters and positions on the input images during training. The translation varied from 0 to 50 mm around the initial position of the needle in the liver and the rotations between 0 and 0.3π. For testing, we rendered the needle as it was inserted into the liver. This training and testing process is specific to the variant denoted 'Ours with visible needle' in Fig. 4.11. With visible needles, the performance of our method is very similar, with a mean dice of 0.805 ± 0.009.

4.5.3.6 Sensitivity analysis

In order to determine common parameters for all datasets, we designed several experiments to explore optimal design choices. These experiments were performed on the porcine dataset. We first explored 4 design choices for the loss function: 1) Mean Squared Error (MSE) loss between the predicted and ground truth displacement field 2) Reprojection MSE loss between the predicted and ground truth displacement field 3) Reprojection MSE loss + 3D soft Dice loss between the predicted and ground truth 3D segmentations 4) Reprojection MSE loss + 2D soft Dice loss between the predicted and ground truth 2D segmentations (see Sec. 4.2.3.2 for loss computation details).

In 5), we performed a hyperparameter search to determine the optimal value of the mixing coefficient λ between the reprojection loss and the 2D soft Dice loss.

We also evaluated the impact of the Backprojection module (described in Sec. 4.2.3.1) over other design choices such as 2D-3D reshape without pooling and 2D-3D reshape with average pooling. These variants are denoted Backproj, NoBackproj, and NoBackproj-AvgPool respectively. In the NoBackproj and NoBackproj-AvgPool variants, the reshape was performed such that the spatial size of features input to the decoder was the same as in the Backproj variant. For the NoBackproj variant, this resulted in 3072 (instead of 768) 3D feature maps, and due to memory constraints, the batch size was reduced from 12 to 3. For the NoBackproj-AvgPool variant, average pooling was performed on the spatial dimensions before reshaping, such that the 3D feature map shape was the same as in the Backproj variant.

Finally, we compared the performances of the network with and without the data augmentation process described in Sec. 4.2.4.

For each case, the network was trained from scratch and the performance of each network was measured at the last epoch.

4.5.3.7 Loss comparison

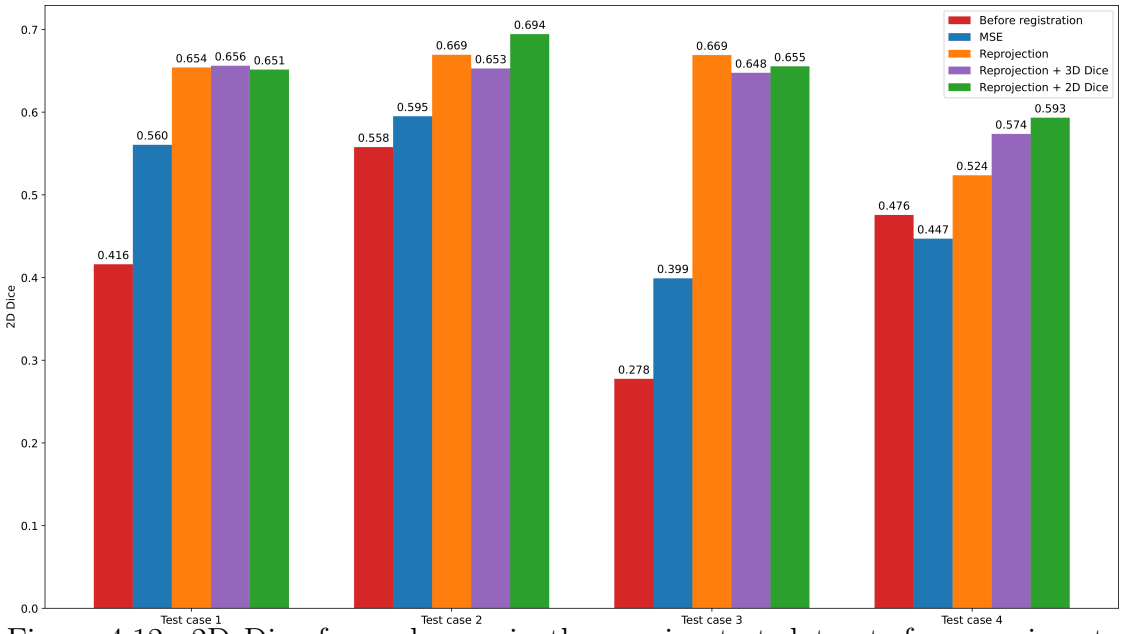


Figure 4.12: 2D Dice for each case in the porcine test dataset, for experiments 1) to 4), where different loss functions are used to train the network. The loss function used in 4) has the best average performances across cases.

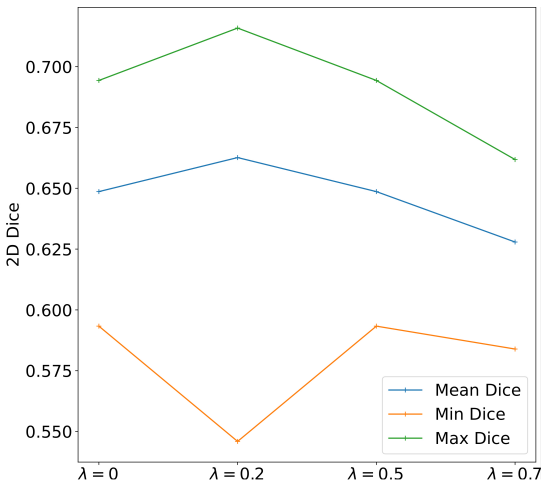


Figure 4.13: Prediction Dice values for all porcine test samples, with varying λ. Compared to other values, $\lambda = 0.5$ is the most robust across cases.

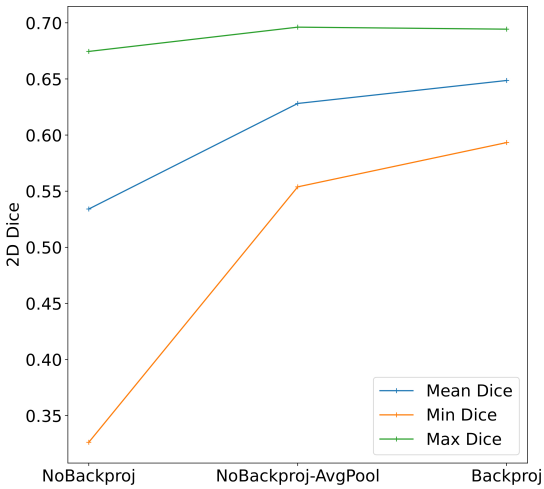


Figure 4.14: Prediction Dice values for all porcine test samples of the NoBackproj, NoBackproj-AvgPool, and Backproj variants. The Backproj variant obtains the best results across variants.

Fig. 4.12 shows the results of experiments 1), 2), 3), and 4) where our proposed loss combination is compared against alternatives. Fig. 4.13 shows the results of experiment 5), where different values of the hyperparameter λ for our loss combination are compared. Based on this experiment, we chose $\lambda = 0.5$ for other experiments, because it was the most robust across cases. Other values of λ demonstrated suboptimal performances for at least one case and were thus rejected.

4.5.3.8 2D-3D translation

To transform 2D feature maps into 3D feature maps, an additional dimension must be introduced. Fig. 4.14 illustrates the performance difference between the three design choices for the 2D to 3D transformation of the feature maps. Fig. 4.15 shows the mean difference between the displacement field predicted for the base input and the displacement fields predicted for 100 perturbed inputs, projected on the input image.

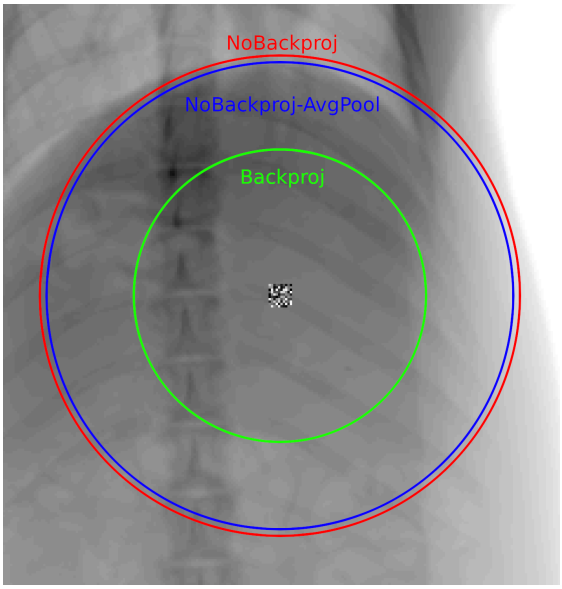


Figure 4.15: In this experiment, the network input is perturbed in a small square with random noise. For each of the NoBackproj, NoBackproj-AvgPool, and Backproj variants, the mean output perturbation is measured and projected on the image. The circles represent the radius in which the output perturbation is at least 0.5 mm. The Backproj variant shows the least spatial extension of the perturbation, demonstrating great spatial correspondence between variations in the input and output.

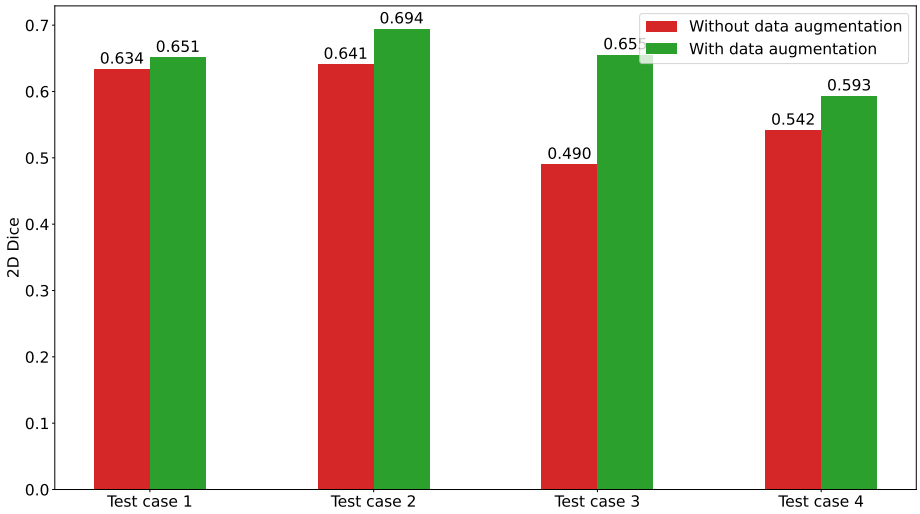


Figure 4.16: 2D Dice with and without data augmentation, for each case in the porcine test dataset. The network always performs better when data augmentation is used.

4.5.3.9 Data augmentation

As introduced in 4.2.4, we used a data augmentation scheme to improve the robustness of the network to changes in image appearances. Fig. 4.16 shows the performance difference between the network trained with and without the proposed data augmentation scheme.

Based on the results of this sensitivity analysis, the parameters we used to measure the performance of our network on the porcine and human datasets are $\mathcal{L} = \mathcal{L}_{\varphi^{2D}} + \lambda \mathcal{L}_{s^{2D}}$ with $\lambda = 0.5$, use of Backprojection to transform 2D feature maps to 3D, and data augmentation to improve network robustness to appearance changes.

4.5.4 Discussion

In this work, we presented a fluoroscopy to CT registration method to recover both breathing and surgically induced deformations. We evaluated our method on a porcine dataset representing common types of deformation in the context of image-guided liver therapy. Table 4.3 demonstrates the potential of our method for contrast-free, fluoroscopy-guided liver therapies. On this dataset, our method was able to recover the shape of the vascular trees for three different types of

intervention-related deformations with good accuracy.

Our results in Fig. 4.10 on the breathing dataset additionally showed that our method is able to recover breathing motion without requiring preoperative 4D-CT acquisition for training data. We obtain comparable performances with a state-of-the-art method, IGCN+, which was trained on patient-specific breathing motion. Our results in Fig. 4.11 show that, thanks to its domain-agnostic training process, our method is able to register a surgically induced deformation with good accuracy, while the IGCN+ method, trained on breathing motion, fails on this unseen deformation. This validates our domain-agnostic approach in the context of fluoroscopy-guided interventions, as it is able to recover intervention-related deformations, which can potentially be of large amplitudes ((Heizmann *et al.*, 2010)). We also demonstrated in Fig. 4.11 that training the network with needles randomly overlaid on the input image makes it robust to the presence of a needle in the image at inference time, as occurs during percutaneous interventions.

We performed a sensitivity analysis in Sec. 4.5.3.6 to justify and evaluate the impact of our design choices. We measured the performance improvement obtained when formulating the segmentation recovery as a training objective using a combination of a 2D reprojection loss and a 2D soft Dice loss. Fig. 4.12 shows that the greatest improvement is obtained when replacing the 3D MSE loss with the 2D reprojection loss. Adding the 2D soft Dice loss brings an additional moderate performance improvement. An explanation for this result is that the 2D Dice loss $\mathcal{L}_{s^{2D}}$ specifically targets the registration accuracy on the vessels, while the loss $\mathcal{L}_{\varphi^{2D}}$ aims to indiscriminately recover the displacement field everywhere, teaching the network the relationship between the 2D input and associated 3D motion. In Fig. 4.13, an hyperparameter search was presented to determine the best mixing coefficient λ between the two losses. Our criterion for hyperparameter selection was that our method should perform comparably across all cases, rather than maximize average performance. Thus, we chose $\lambda = 0.5$ as this value offered the best compromise between average performance and robustness across cases. Fig. 4.16 validates the effectiveness of the data augmentation method proposed in (Grimm *et al.*, 2021) for our application, with important performance improvements on some cases. We hypothesize that slight changes in image appearance due to inter-fractional anatomical variations, caused, for example, by the digestive process, may induce prediction errors that are mitigated by this data augmentation approach.

Finally, we evaluated whether our back projection layer was useful to translate 2D features into 3D features. A natural choice would be to use the C channels of the feature maps as a depth dimension. However, Fig. 4.14 clearly illustrates that this approach is too naive and leads to performance degradation. To respect the projective aspect of the 2D-3D transformation, the Backproject module introduced

in Sec. 4.2.3.1 transforms the 2D feature maps using the projection matrix. This introduces a direct correspondence between a position $\mathbf{x}$ in the volume and a pixel $\mathbf{u} = \mathbf{P}\mathbf{x}$, which can be observed by randomly perturbing the input at a position $\mathbf{u}$ and measuring the change in the output displacement field. In Fig. 4.15, only the Backproj variant displays a local perturbation of the displacement field around the perturbed input, while the NoBackproj and NoBackproj-AvgPool variants display global perturbations of the displacement field. This indicates that the Backproj variant indeed incorporates a 2D-3D spatial correspondence.

While we used the Dice coefficient to evaluate our method, care must be taken to interpret the value of the Dice coefficient of thin structures such as blood vessels. For example, a Dice of 0.6 may seem low when compared with other values in the state of the art, but this value is highly influenced by the small size of the vessels and the differences in the acquisition of the baseline and deformed vessel trees. Notably, the lengths of corresponding branches in the baseline and deformed vessel trees are different, which affects the Dice coefficient. It is, in turn, more informative to look at the relative change before and after registration to evaluate results. Fig. 4.17 shows how the value of the 2D Dice coefficient changes when the intrahepatic vessels and the liver are translated. Due to their long and thin shape, the Dice metric varies far more rapidly for the intrahepatic vessels than the liver. This is showcased in the top right part of Fig. 4.17, which shows how a small shift (5mm translation) affects alignment for the liver and intrahepatic vessels.

Even though our results show that our method has the potential to enable augmented fluoroscopy-guided interventions, challenges remain before it can be used in the operating room. First, in this study, we did not validate the effectiveness of our method on real fluoroscopic images, as no paired fluoroscopy/CT datasets were readily available to test our deformable registration method. Obtaining a fluoroscopic image with a perfectly corresponding CT scan volume is not easily done due to practical constraints, but it could be the subject of future work with the use of a robotized CBCT device. However, other works focusing on rigid registration problems demonstrated that specialized synthetic training techniques could be used to bridge the domain gap between DRR images and fluoroscopic images (Grimm *et al.*, 2021; Gao, Killeen, *et al.*, 2023; Jaganathan *et al.*, 2023) reporting clinically relevant performances. In our case, using the data augmentation technique in (Grimm *et al.*, 2021), originally developed to bridge this domain gap, improved performances even though DRRs were used as testing data.

Additionally, while our method is suitable for augmenting fluoroscopic images in real-time, it is still limited by the weak depth information in the fluoroscopy. To remediate this, a biomechanical model of the organ (in this case the liver) could be used to generate physically accurate training data. The biomechanical model could also be used in the loss to guide the prediction of the network towards

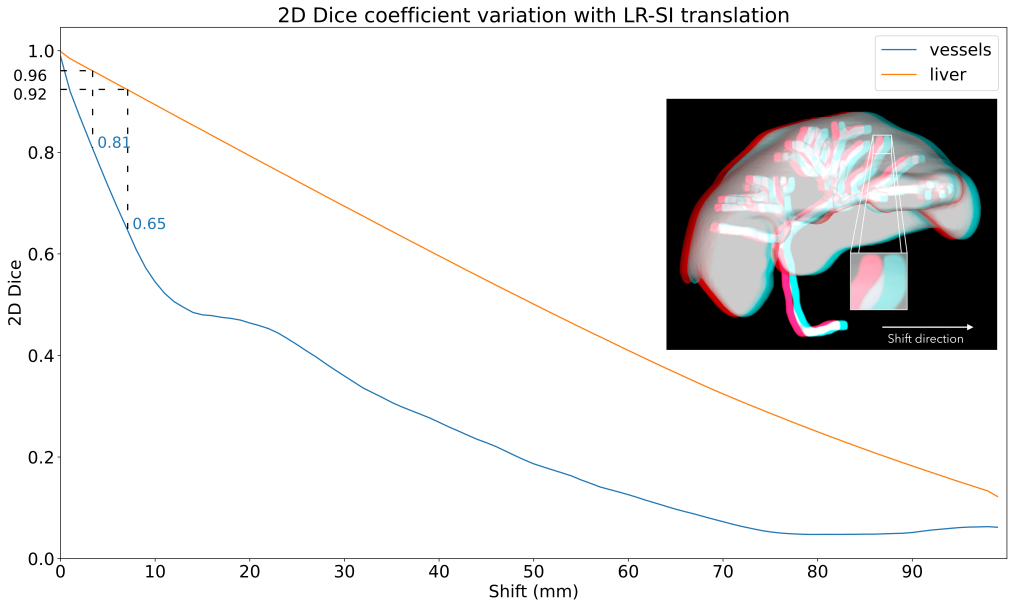


Figure 4.17: This figure illustrates how the Dice coefficient evolves when 2 shapes, initially perfectly overlapped, are progressively misaligned. We particularly emphasize how the Dice value evolves as a function of the shift when the shape is a relatively large structure (e.g. an organ) and when the shape is a thin structure (e.g. blood vessels). This puts our results into perspective: a Dice value of 0.65 (obtained on the porcine dataset) would correspond to a Dice of 0.92 on a whole organ, while a Dice value of 0.81 (obtained on the breathing motion dataset) would correspond to a Dice of 0.96 on the liver.

physically plausible displacements.

Finally, while we used a soft Dice loss in our loss function, application-specific losses, such as the Centerline Boundary Dice Loss (Shi *et al.*, 2024) specifically designed for vascular structures, could be used for potentially improved performances.

Moving forward, our focus will be allocated to refine the method's accuracy by incorporating a physics-based deformation model. Additionally, we plan to evaluate its performance using real fluoroscopic images, aiming to validate and enhance its practical applicability in diverse clinical settings.

4.6 Fluoroscopy-guided autonomous guidewire navigation

4.6.1 Introduction

Navigating a catheter and guidewire through the vascular system in a safe and efficient manner is crucial to minimize the patient's and clinician's exposure to X-ray radiation from the fluoroscopic imaging system. This task demands a thorough understanding of the anatomy, superior device control, and a comprehensive grasp of fluoroscopic visualization. However, even seasoned clinicians may take considerable time to reach specific targets. Robotic systems can potentially enhance this process (Puschel *et al.*, 2022). Yet, these robots are still master-follower systems that operate the devices based on the clinician's inputs. To further assist the clinician, current research is shifting towards the creation of autonomous and semi-autonomous systems. Among the semi-autonomous systems, Zhang *et al.* (J. Zhang *et al.*, 2024) proposed an algorithm to maintain the tip of a robotized bronchoscope at the center of the airways. This AI-based algorithm uses both bronchoscopic images and human commands as inputs and predicts a corrective motion. Autonomous systems generally rely on Deep Reinforcement Learning (DRL) and use fluoroscopic images to predict a control action (rotation and translation) to be executed at the device's proximal end. Some research trains and applies the learned control entirely in simulated environments (W. Tian *et al.*, 2023). Other studies train the neural controller using images of the phantom where the navigation will later be performed (S. Wang *et al.*, 2022; Kweon *et al.*, 2021) while others perform the training in a simulated environment and then use images of the phantom during navigation (Karstensen, Behr, *et al.*, 2020).

The limitations of current research are two-fold. Using fluoroscopic images as input to the neural network can cause uncertainties about the orientation of the tip, leading to prediction errors. Also, the training process does not generalize well and requires individual training for each patient. As reported in (Miranda *et al.*, 2023; Kirk *et al.*, 2023), learning controllers that can perform tasks in both familiar and unfamiliar environments remains a significant challenge in DRL.

To tackle this problem in the context of endovascular procedures, Kweon *et al.* (Kweon *et al.*, 2021) suggested a segment-wise learning method to speed up training using human demonstrations, transfer learning, and weight initialization. However, this method still necessitates network training each time the environment is altered or expanded. Similarly, in the research conducted by Karstensen *et al.* (Karstensen, Ritter, *et al.*, 2023), the controller performance dropped from a 75% success rate in navigating known anatomies to 29% when real patient vessels were used. Chi *et al.* (Chi, J. Liu, *et al.*, 2018; Chi, Dagnino, *et al.*, 2020) proposed

different strategies to obtain an optimal control of the device. In (Chi, Dagnino, *et al.*, 2020), they used a generative adversarial imitation learning method aimed at learning the catheterization of different arteries, with a success rate of about 70% when the aortic type was altered. In (Chi, J. Liu, *et al.*, 2018), they trained a statistical model to perform the cannulation of the innominate aorta and applied the same controller to variations of the aortic arch type. This technique reported an average 98% success rate in new but very similar geometries, using human demonstrations for each new task.

Recently, we proposed a DRL method that achieves excellent generalization thanks to a specific training strategy (Scarponi, Duprez, *et al.*, 2024). Using a set of only 4 bifurcation shapes, and a shape-invariant observation space, the learned controller was able to navigate complex, unseen anatomies. Three main assumptions were made: the vessels have a nearly constant radius, the bifurcations always have 2 exit vessels, and the anatomy is not moving or deforming during navigation. The first assumption is not a limitation of the method but a consequence of using a unique guidewire during the navigation. With a constant tip shape, only vessels of a compatible diameter can be accessed. Branching patterns with one entry vessel and two exit vessels were chosen as bifurcation is the most common pattern (Singh *et al.*, 2017).

This study addresses the third assumption (static anatomy) and proposes two main contributions: a training strategy able to learn a control of the device even when the vascular anatomy is moving and/or deforming (see Sec. 4.6.2.1), and a method to estimate the motion of the anatomy from single view fluoroscopy images (see Sec. 4.6.2.2). The combination of these two contributions makes it possible to automatically navigate across a moving anatomy under fluoroscopic imaging, even without injecting a contrast agent. Our results (see Sec. 4.6.3) illustrate the genericity of the training, and the excellent performance of our method, even when applied to complex, deforming anatomies only observed through 2D fluoroscopic imaging.

4.6.2 Guidewire control in dynamic environments

4.6.2.1 Learning to navigate dynamic environments

Our objective is to develop a generalized neural controller able to control the motion, in particular the rotation, of a guidewire through a complex, deforming vascular tree, from its insertion point until a given target is reached (see Fig. 4.22, 4.24), while it is advanced at a variable speed. This control is performed at the proximal end of the device, and accounts for both the device and anatomy deformations during navigation. Our learning method relies on five main elements: 1) an efficient DRL algorithm; 2) a fast and accurate simulated environment that can be

updated based on external input; 3) an observation space robust to affine transformations of the anatomy; 4) a specific reward function; 5) an optimal choice of training anatomies. They are described below.

Training algorithm Reinforcement Learning (RL) constitutes one of the areas of machine learning. In this specific branch, an agent learns to achieve specific goals by interacting through its actions with an environment. The problem is usually formulated as a Markov Decision Process (Bellman, 1957), *i.e.* $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, in which $\gamma \in [0, 1]$ defines the discount factor, and $\mathcal{S}$ and $\mathcal{A}$ represent a set of states and actions respectively. Each action $a_t \in \mathcal{A}$ induces a transition in the system from the current state $s_t \in \mathcal{S}$ to the next state $s_{t+1} \in \mathcal{S}$, and is chosen by a policy π, mapping states to actions $\mathcal{S} \to \mathcal{A}$. The probability density of the next state $s_{t+1} \in \mathcal{S}$ given the current state $s_t \in \mathcal{S}$ and action $a_t \in \mathcal{A}$ is denoted by $P(s_{t+1}|s_t, a_t)$. For each transition, the agent receives a reward $r(a_t, s_t)$. The agent observes the environment through the observation space Ω, which constitutes a total or partial description of the environment itself.

As in our previous work (Scarponi, Duprez, *et al.*, 2024), we adopted in this study the Soft Actor-Critic (SAC) algorithm, which outperformed previous algorithms (Haarnoja *et al.*, 2018) such as the deep deterministic policy gradient (DDPG), largely used for autonomous catheter navigation (W. Tian *et al.*, 2023; Karstensen, Behr, *et al.*, 2020). In Eqn. (4.6) the objective function of the SAC algorithm is reported, in which the entropy term $\mathcal{H}(\pi(\cdot|s_t))$ is introduced. This term, which constitutes the main novelty of SAC algorithm, promotes the exploration of the environment and discourages the repetition of actions that may exploit inconsistencies in the approximated Q-function.

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t, a_t) \sim \rho_\pi} [r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t))]. \tag{4.6}$$

Here $\rho_\pi(s_t)$ and $\rho_\pi(s_t, a_t)$ represent respectively the state and state-action marginals of the trajectory distribution induced by a policy $\pi(a_t|s_t)$.

Simulation of the training environment The virtual environment used to train the DRL algorithm is based on a physics-based simulation of the device and its interactions with the vessel walls. We developed our simulator using the open-source SOFA framework (Faure *et al.*, 2012) and relying on Timoshenko beam theory (Bitar *et al.*, 2015) to model the physics of the guidewire. The system to be solved is reported in Equation (4.7) in its matrix form.

$$(\mathbf{M} - dt^2 \mathbf{K}) \Delta v = dt \cdot f(x(t)) + dt^2 \cdot \mathbf{K} v(t), \tag{4.7}$$

where $\mathbf{M}$ represents the mass matrix, $\mathbf{K}$ the stiffness matrix and dt the time step. v and Δv denote the velocity and the velocity variation respectively and f, which is a function of the current positions $x(t)$, represents the internal and external forces applied to the system.
The interactions between the vessel wall and the guidewire are computed using a constraint-based approach and the position of the vascular anatomy is updated at each time step based on the current fluoroscopic image (see Sec. 4.6.2.2). Equation (4.7) then becomes:

$$(\mathbf{M} + dt\frac{df}{d\dot{x}} + dt^2\frac{df}{dx})\Delta v = -dt(f + dt\frac{df}{dx}v) + dt\mathbf{H}^T\lambda, \tag{4.8}$$

where $\mathbf{H}^T\lambda$ is the vector of constraint forces, with $\mathbf{H}$ containing the constraint directions arising from the collision detection, and λ the Lagrange multipliers. The physics of the guidewire model is then corrected by computing the contact force λ using a Gauss-Seidel algorithm (Jourdan *et al.*, 1998).

Using a Block Tridiagonal solver, the navigation of the virtual device is simulated at 90 frames per second, maintaining both short training times (Sec. 4.6.3) and a sufficient level of accuracy (2.0 ± 0.9 mm error between the simulated guidewire and the shape of a scanned guidewire inserted inside a vascular phantom).

Reinforcement learning strategy As illustrated in Scarponi *et al.* (Scarponi, Duprez, *et al.*, 2024), the definition of the observation space, and choice of the training geometries, are essential to learn a generalizable control. In this work, we keep a similar strategy: we train the RL algorithm on a set of bifurcation patterns, unrelated to the test anatomies. This local vascular shape is represented by both a surface mesh and a centerline. The only assumption about the training shapes is that the diameter of the vessels is nearly constant and that they have a Y-shaped topology.

We then extend the work from (Scarponi, Duprez, *et al.*, 2024) in two areas. First, we augment the training database by introducing shape variations of the training anatomy during the training process (i.e. similar to sim-to-real approaches). This shape variation is continuous throughout space and time, to avoid discontinuities in the displacement field that would cause errors in the simulation. Second, we formalize the shape generation process by making it procedural, rather than handcrafted. We characterize the 3D vessel shape from its centerline C from which it is extruded. C is defined as $C(\phi_i, \nu_j)$ with $i \in \{1, 2\}$ and $j \in \{1, .., 6\}$, where ϕ_1 and ϕ_2 define the angles between the bifurcation branches and ν_j are the tangent of the centerline shape at each endpoint (see Fig. 4.18). Starting from the simplest geometry (a Y-shaped bifurcation with straight branches) we progressively deform this shape into a series of other shapes, by varying smoothly ϕ_i and

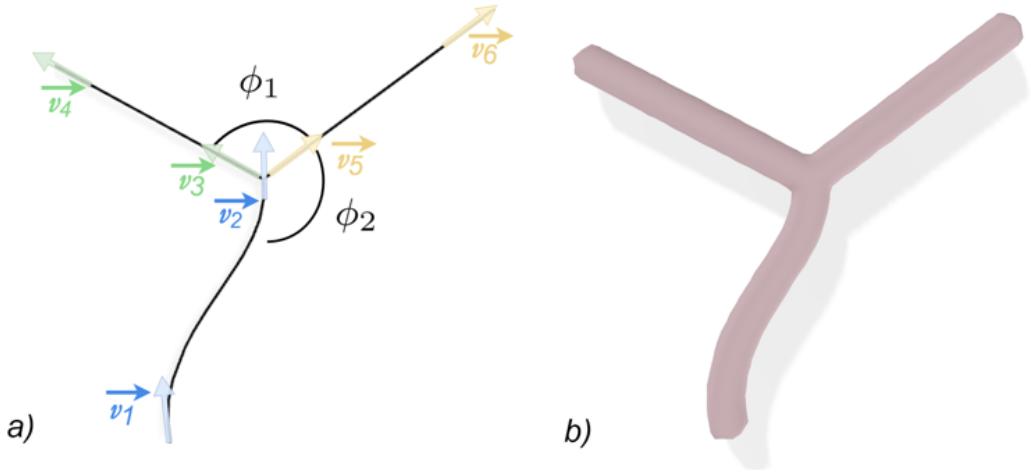


Figure 4.18: Procedural training shape generation process. a) Centerline $C(\phi_i, \nu_j)$ of the vessel with $i \in \{1, 2\}$ and $j \in \{1, .., 6\}$. ϕ_1 and ϕ_2 are the angles between the bifurcation branches and ν_j are the tangents to the centerline shape at each endpoint. b) 3D shape of the vessel obtained by extrusion of the centerlines C.

ν_j and maintaining a constant vessel diameter. Fig. 4.19 illustrates this process. Let's call $\mathcal{B}$ the set of all the bifurcation shapes we generate through our process. We split $\mathcal{B}$ into a series of N subsets $\mathcal{B}_k$ of random length, such that $\cup_{k=1}^{N} \mathcal{B}_k = \mathcal{B}$. Each subset $\mathcal{B}_k$ represents a different range of shape variations, from small deformations to large ones. These shape variations are then used as training anatomies during the learning processes of our neural controller. For each training episode, a target is randomly selected, as well as a subset $\mathcal{B}_k$ of the varying training anatomy. The initial rotation of the guidewire around its axis and its orientation relative to the centerline are chosen randomly to enhance the exploration of the environment and, during each episode, the velocity of the device is also randomly modified.

Nearly shape-invariant observation space To enforce generalization of the learned control, we proposed as in (Scarponi, Duprez, *et al.*, 2024) an observation space that is rotation and translation invariant, but also shows little sensitivity to the shape variation of the bifurcation. This is achieved by defining observations that are relative to the position of the device in the environment. In this work, we expand the observation space by adding elements that permit to navigate geometries that are different both in shape and size, with the sole caution of using a guidewire compatible with the vessel diameter. The observation space Ω is constructed as follows:

$$\Omega = \{\boldsymbol{\zeta}_t, \boldsymbol{\zeta}_{t-ndt}, \lambda_t, \lambda_{t-ndt}, a_t, \omega, d_v\}$$

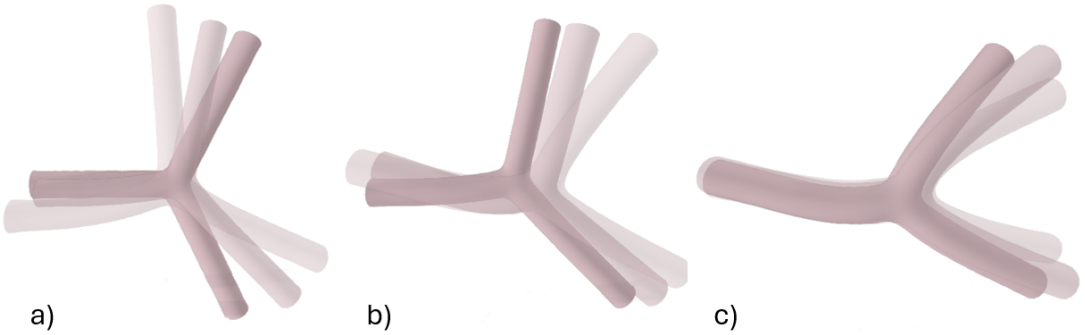


Figure 4.19: Examples of $\mathcal{B}_k$ subsets. Each subset $\mathcal{B}_k$ represents a range of bifurcation shape variation and $\cup_{k=1}^{N}\mathcal{B}_k = \mathcal{B}$, where $\mathcal{B}$ defines the complete set of bifurcation shapes considered in this work.

- Let $\mathbf{t}_i, i \in \{1, \ldots, N\}$ be the tangent vector at the coordinate $\mathbf{x}_i$ along the tip of the guidewire, and $\mathbf{c}_j, j \in \{1, \ldots, N\}$ the tangent vector of the centerline at position $\mathbf{x}_j = \mathbf{x}_i + \mathbf{h}$. We define $\zeta_i = \mathbf{t_i} \cdot \mathbf{c_i} \ \forall \ i \in \{1, \ldots, N\}$. To handle dynamic environments, $\mathbf{c_i}$ must be updated. This does not require changing the observation space defined in (Scarponi, Duprez, *et al.*, 2024) but necessitates estimating this change from live images during an intervention. Our method for handling this challenge is described in Sec. 4.6.2.2.

- We then define $\boldsymbol{\zeta}_m = [\zeta_1, \zeta_2, \ldots, \zeta_N]_m$, with $m \in \{t; t - ndt\}$.

- λ_t and λ_{t-ndt} represent the distance between the tip of the guidewire and the target at time t and $t - ndt$, normalized with respect to the initial distance to the target λ_0. λ_0 is defined as the target distance at the entrance of the bifurcation region.

- a_t is the action that determines the transition of the system from s_{t-ndt} to s_t.

- $\omega = \mathbf{k}_p \cdot \mathbf{w}_p$, where $\mathbf{k}_p$ and $\mathbf{w}_p$ are the projections of the vectors $\mathbf{k}$ and $\mathbf{w}$ onto a plane Γ perpendicular to the centerline of the branch leading to the target (see Fig. 4.20b). $\mathbf{k}$ represents the radial vector of curvature located in the middle of the curved tip, and $\mathbf{w}$ is the vector describing the direction of the wrong branch. To be robust to different vessel dimensions, both in terms of vessel diameter and exit branch length, the proper choice of $\mathbf{w}$ and Γ normal vector (n_Γ) is crucial. $\mathbf{w}$ norm is proportional to the vessel diameter and its starting point is fixed at the center of the bifurcation, while n_Γ magnitude is proportional to the squared diameter of the vessel and it originates from the projection of the guidewire distal end onto the centerline (see Fig. 4.20b).

- $d_v = \mathbf{v} \cdot \mathbf{c}$, where $\mathbf{v}$ describes the current velocity of the guidewire and $\mathbf{c}$ the tangent to the centerline near the tip of the guidewire (see Fig. 4.20c).

It is important to notice that all the parameters used to build the observation space can be computed in both the virtual (training) environment and in a real setup. The vessel geometry can be retrieved from preoperative images and updated intraoperatively (see Sec. 4.6.2.2) and the tip shape of the guidewire can be reconstructed from Fiber Bragg Gratings (FBG) data (Al-Ahmad *et al.*, 2020) using an optical fiber embedded in the catheter or guidewire.

Reward function Another key element to learning the optimal action is the engineering of the reward function. We design our reward function as the weighted sum of three terms:

$$r(s_t, a_t) = \underbrace{\frac{2}{1 + e^{5(\omega - 0.1)}} - 1}_{a} + \underbrace{0.5\,(1 - \lambda_t)}_{b} + \underbrace{(-0.2|a_t|)}_{c},$$

where part a of the reward function encourages the agent to obtain a tip direction $\mathbf{k}_p$ opposite to $\mathbf{w}_p$ (see Fig. 4.20). This function is a modified version of the sigmoid activation function. The output of part a is a decreasing function taking its values in $[-1; 1] \in \mathbb{R}$. Part b of the reward increases as the target is approached, while part c discourages the agent from rotating the instrument when it is unnecessary.

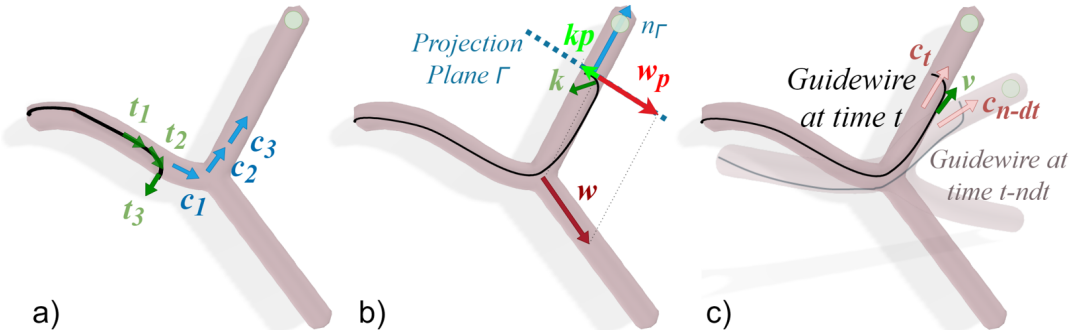


Figure 4.20: Observation space: 1) dot product between $\mathbf{t}_i$ (tangents to the tip of the device) and $\mathbf{c}_j$, describing the downstream centerline, with $i \in \{1, 2, 3\}$ and $j \in \{1, 2, 3\}$ (a), 2) normalized distance between the tip of the guidewire and the target, 3) chosen action, 4) dot product between $\mathbf{k}_p$, describing tip's direction, and $\mathbf{w}_p$, describing the direction of the branch that does not lead to the target (b) 5) dot product between $\mathbf{v}$, describing the velocity of the guidewire and $\mathbf{c}$, describing the centerline (c).

4.6.2.2 3D vascular motion estimation from fluoroscopic images

The observation space Ω described previously includes the relative position of the device tip with respect to the vessel centerline and target, among other things. When the shape and position of the vessels are changing (e.g. due to cardiac or respiratory motion) we must update this information such that the neural controller can perform optimally.

To recover this motion, it is necessary to use a real-time imaging modality that presents sufficient contrast between the vessels and the surrounding tissue. Fluoroscopy is the only imaging modality that meets these criteria and is currently used in the vast majority of endovascular interventions. However, it requires the injection of a contrast agent to be able to visualize vessels in the image, and it provides only a two-dimensional image.

Various methods have been devised to overcome this limitation and recover 3D motion from a single fluoroscopy, mostly in the context of free-breathing radiotherapy (Wei *et al.*, 2020; Nakao *et al.*, 2022; Shao, Y. Li, *et al.*, 2023). While these methods demonstrate clinically relevant target localization accuracy, their use of a statistical motion model to generate training data limits them to recovering predetermined motion patterns, which can restrict their clinical applicability. Moreover, the accuracy of these methods to recover the shape of vessels has not yet been evaluated.

Fluoroscopy-based vessel motion prediction In this work, the network architecture presents the following differences with the architecture detailed in Sec. 4.2.3:

- In this work, an MSE between the ground truth displacement field and the predicted displacement is used to optimize the network parameters, with the displacement in the direction perpendicular to the projection set to 0, and no learning rate scheduler is used.
- A Cosine Annealing learning rate scheduler (Loshchilov *et al.*, 2016), which decays the learning rate following a cosine schedule, was used.
- 6 encoder and decoder layers are used instead of 10
- The shape of the predicted and ground truth displacement fields is $(64, 32, 64)$ instead of $(128, 64, 128)$.
- The domain randomization data augmentation described in Sec. 4.2.4 was not used.

The data generation process is identical to the data generation process described in Sec. 4.2.2.

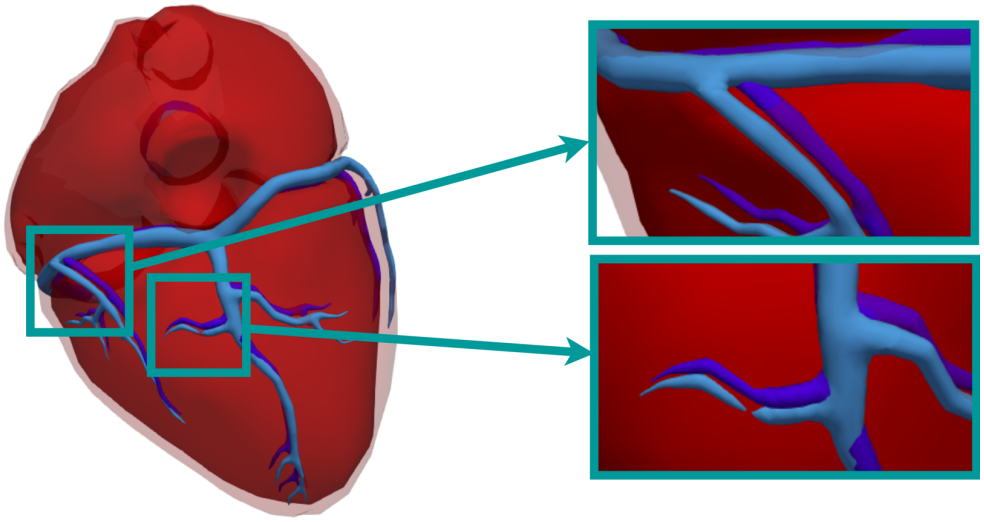


Figure 4.21: Visualization of the coronary motion during a cardiac cycle. The heart volume varies by about 12% in 1 second during a cardiac cycle.

4.6.3 Results

We illustrate the performance of our neural controller and fluoroscopy-based motion estimation in two examples. The first one is a typical example of endovascular navigation, where a guidewire is advanced in the coronary arteries of a beating heart. In the second example, we show a scenario where a guidewire is advanced through the venous system of the liver, as done during the diagnosis and treatment of portal hypertension. Based on the method described in Sec. 4.6.2.1 and using Stable Baselines3's SAC implementation (Raffin *et al.*, 2021), we train the neural controller using a learning rate of 10^{-4}, a buffer size of 10,000 and a batch size of 256. The discount factor is set to 0.98 and the entropy coefficient is learned during the training. The actor and the critic networks are composed of three 256-neuron layers and the model, updated at every time step, is trained for 175,000 time steps, with a dt of $0.01\,s$. The whole training process only requires 6 hours of computation on an Intel(R) Core(TM) i7-13700KF processor with 32 GB of RAM. The training anatomies are generated as explained in Sec. 4.6.2.1 with a constant vessel diameter of $4\,mm$. A suitable guidewire is used to navigate the anatomies, with a $4.5\,mm$ long tip and a tip curvature of $0.38\,mm^{-1}$. For each anatomy and test case, the controller is evaluated on a total of 100 episodes, where an episode is defined as the navigation from the insertion point to the target location. An episode is considered successful if the guidewire, steered by the controller, reaches the target location. Four and five distinct target locations were chosen for respectively the heart and the liver, each involving the navigation of a minimum of 2

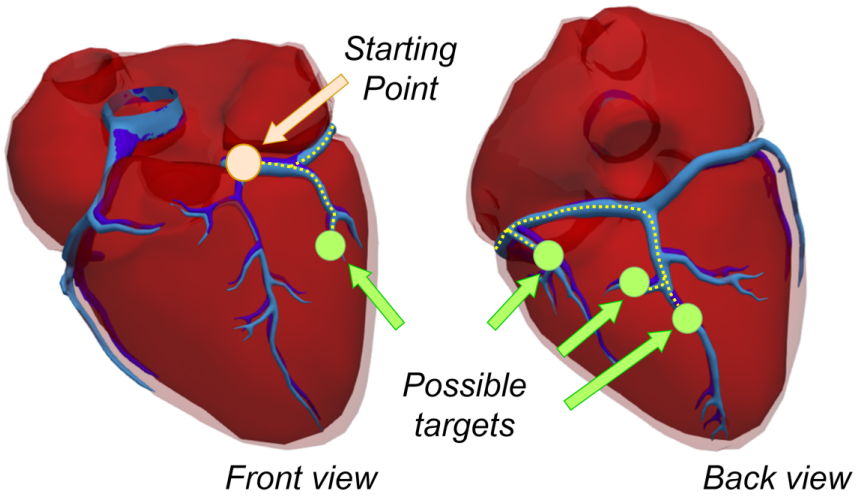


Figure 4.22: Illustration of the different paths and targets the neural controller needs to navigate. The insertion point is shown in orange, and the randomly selected targets are shown in green. An example of a path is depicted as a dotted line.

and a maximum of 4 bifurcations. For each test episode, a random target and a random starting rotation of the guidewire were chosen. We report in Table 4.4 the percentage of successful episodes for each test.

4.6.3.1 Navigation in coronary arteries during cardiac motion

In this example, we demonstrate the ability of our neural controller to navigate a dynamic environment without prior training on either this anatomy or this particular deformation (see Fig 4.21). The heart model was reconstructed from Magnetic Resonance imaging data but the motion was generated synthetically, allowing to know the shape of the vascular tree and the centerline position at each time step. In this case, since the dimensions of the anatomy are similar to the dimensions of the training geometries, with an entry diameter of $3.8\,mm$, the same guidewire used during the training is adopted. The efficacy of the controller is tested in three different scenarios: a static case, in which the anatomy is not moving, a dynamic case in which the heart is beating, but the location of the centerline is not updated and a dynamic case in which the heart is beating and the position of the centerline is updated. For this anatomy, we selected 4 targets shown in Fig. 4.22 and randomly chose one for each of the 100 test episodes. In all test cases, our controller proves its ability to navigate the coronaries both in static (90% success rate, Table 4.4 a) and dynamic (97% success rate, Table 4.4 c) conditions, maintaining its performance also when navigating the anatomy without any knowledge

about vessel deformation (89% success rate, Table 4.4 b).

4.6.3.2 Navigation in hepatic veins during respiratory motion

In this section, we focus on a different clinical context, such as the endovascular treatment of hepatic venous outflow obstruction (Ghibes *et al.*, 2023) or the endovascular treatment of portal hypertension (Golowa *et al.*, 2012). The key difference when compared to the previous scenario is that, in this case, the vascular tree motion caused by the breathing of the subject is unknown, engaging ourselves in a true clinical scenario, in which the moving anatomy is only visible in fluoroscopic images, in 2D. Using the neural network described in Sec. 4.6.2.2, we estimate, in real-time, the 3D position of the vessels' centerlines. Using a patient's abdominal CT scan, a training dataset, composed of 18,000 samples, was generated and used to train the motion prediction neural network, as described in Sec. 4.2.2. The test dataset contains a series of fluoroscopic images covering 5 inhale/exhale periods for a total of 50 samples. The main direction of motion is along the Inferior-Superior (IS) and Antero-posterior (AP) axes, with a small motion in the Left-Right (LR) direction, and a sliding motion of the organs against the thoracic cage. The trained network was evaluated using the Target Registration Error (TRE) on the hepatic veins centerlines and the hepatic veins mesh. Across the testing dataset, the mean displacement was $8.74 \pm 4.06\,mm$ and $8.66 \pm 4.05\,mm$ while the mean TRE was $3.74 \pm 2.33\,mm$ and $3.84 \pm 2.37\,mm$ for the centerlines and the hepatic veins respectively. This error is not similar in each direction, since the motion along the direction perpendicular to the image plane is more difficult to estimate compared to the other two directions. This is reflected in the error of the network, which was, on average, below $2\,mm$ for the IS and LR directions and, on average, below $2.6\,mm$ for the AP direction.

We report in Table 4.4 the success rate of our controller, in similar conditions as presented for the heart: a static case, a dynamic case without centerline update, and a dynamic case in which the updated position of the anatomy is reconstructed from non-contrasted fluoroscopic images, thanks to our neural network. Given the dimension of the anatomy, a different guidewire, with a $6.5\,mm$-long tip and a tip curvature of $0.26\,mm^{-1}$ is used. For this second test anatomy, we chose 5 different target locations, as shown in Fig. 4.24, and randomly selected one for each of the 100 test episodes. In this more complex context, our controller demonstrates its efficacy with an 89% success rate (Table 4.4 d) in the static scenario, which is maintained when transitioning to dynamic conditions. In this case, the agent reports a success rate of 93% (Table 4.4 f), while it shows an important performance drop (24% success rate, Table 4.4 e) when trying to navigate the dynamic anatomy without any information regarding the movement of the vessels (no centerline update in the observation space Ω). This shows the significance of our 3D vascular

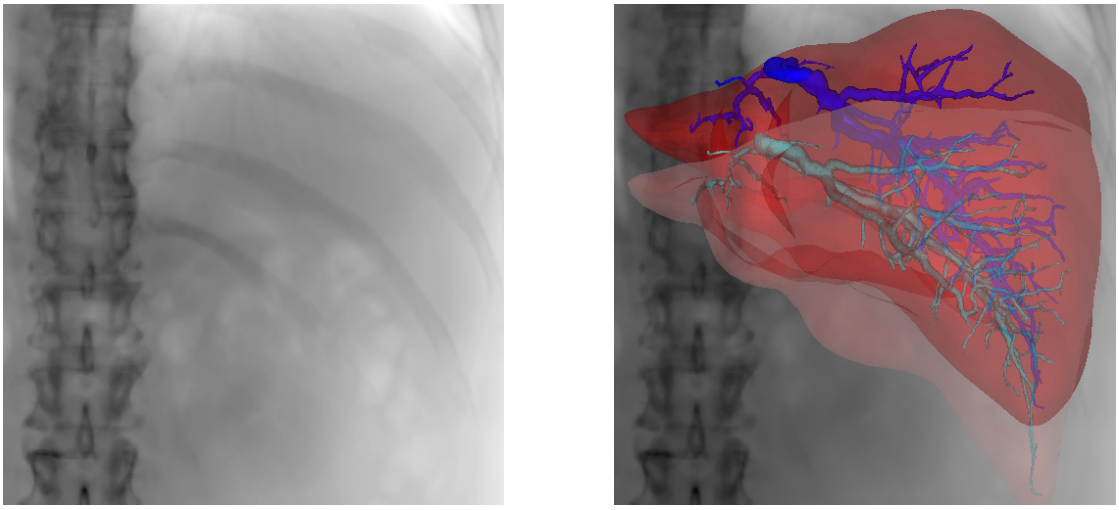


Figure 4.23: **Left**: fluoroscopic image seen by the neural network. **Right**: preoperative position and shape of the liver and its venous system (opaque colors) and prediction of the 3D shape (in semi-transparent color) of both the liver shape and its vascular tree. The centerlines of the veins are also predicted, in real-time, and used by the neural controller.

motion estimation.

4.6.4 Discussion

As illustrated by the results in Table 4.4, our new controller demonstrates its ability to navigate new complex anatomies, composed of various subsequent bifurcations, different in shapes and dimensions, with a mean success rate of 95% in the dynamic anatomies. This is very significant compared to the probability of reaching the designated targets when taking random actions which would lead to an average success rate of 15% for the two scenarios we have considered. This value can be obtained by considering the set $\boldsymbol{Z} = \{z_1, z_2, \ldots, z_N\}$ of N different targets, each requiring the successful navigation of a number, b_{z_i}, of bifurcations. The mean probability of successfully reaching the target is equal to $\frac{1}{N}\sum_{j=1}^{N} 1/2^{b_{z_j}}$.

We compared our new controller with the agent described in our previous work (Scarponi, Duprez, *et al.*, 2024), trained for 150,000 time steps of 0.01 s on anatomies with a $4\,mm$ diameter, consistent with the diameter of the training anatomies used in this work, allowing to use the same guidewire. The new controller outperforms our previous version in terms of both robustness to vessel motion (Table 4.4 c) and robustness to variations in vessel dimensions (Table 4.4, *Liver*). The important performance loss shown by (Scarponi, Duprez, *et al.*, 2024) when navigating the liver anatomy, can be explained by the difference in

Table 4.4: Navigation results summary, in heart and liver.

Heart			
Conditions	Success rate [%]		
	(Scarponi, Duprez, *et al.*, 2024)	Our method	Random
Static[a]	100%	90%	15.6%
Dynamic, no centerline update[b]	68%	89%	
Dynamic, centerline update[c]	82%	**97%**	
Liver			
Conditions	Success rate [%]		
	(Scarponi, Duprez, *et al.*, 2024)	Our method	Random
Static[d]	36%	89%	15%
Dynamic, no centerline update[e]	50%	24%	
Dynamic, centerline updated with our NN prediction[f]	50%	**93%**	

the dimension of the liver geometry, whose entry vessel presents a diameter of $7\,mm$, which almost doubles the diameter of the training anatomies. The high success rate obtained by our controller in the liver's static conditions (Table 4.4 d) demonstrates the adaptability of our new training strategy to anatomies presenting various bifurcation shapes and dimensions. However, when navigating a dynamic environment, the performances of the controller critically drop if the centerlines are not updated (see Table 4.4 e). Our neural network allows computing the new position of the vessels, thus reducing the difference between the real anatomy and the anatomy observed by the controller. In these conditions, our controller demonstrates its ability to navigate various dynamic anatomies both when a synthetic movement is generated (Table 4.4 c) and when the anatomy moves following real vessel movements (Table 4.4 f).

4.6.5 Conclusion

In this study, we presented a neural controller, based on a deep reinforcement learning approach, able to navigate a guidewire in complex, unseen, moving anatomies with various dimensions. In addition, we proposed a method for estimating the 3D motion of the anatomy from single-view fluoroscopy images, even without the injection of a contrast agent. The combination of these two contributions makes it possible to automatically perform endovascular navigation in close to real-world conditions, as illustrated in two scenarios: a beating heart and a liver deformed

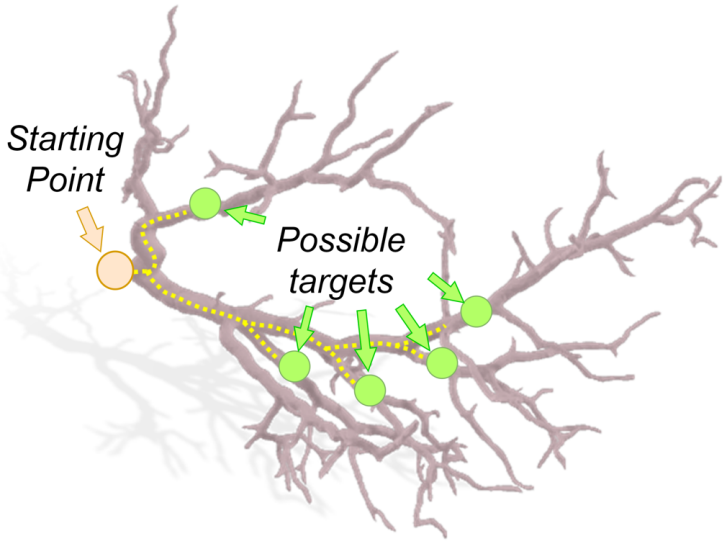


Figure 4.24: Liver venous system with 5 different targets. The controller has to navigate the moving anatomy from the insertion point (in orange) until the designated target, randomly chosen among the possible targets, in green.

under breathing motion. Our method makes it possible to reach random targets within these anatomies with an average success rate of 95%. To the best of our knowledge, this has never been achieved before.

Although our method already accounts for certain real-world conditions (use of actual anatomies generated from patient data, motion estimation from fluoroscopy), a natural future development of this work will consist in testing the neural controller in a vascular phantom. This will require the use of an FBG-based shape sensing method to reconstruct the shape of the guidewire, and an access to an endovascular robot to apply the action taken by the controller to the device. We have already started working on the shape reconstruction from FBG data, and we will first assess the robustness of our controller in a rigid phantom.

4.7 Conclusion

Across the above studies, we thoroughly evaluated our domain-agnostic 2D-3D deformable registration framework. These works show that real-time 2D-3D registration in fluoroscopy-guided interventions is feasible, without requiring fiducial markers or preoperative statistical deformation models.

In our first study (section 4.3), we demonstrated that domain-agnostic data generation enable deformation recovery more robustly than PCA-based data gen-

eration. The advantage of this approach is that it requires only routinely acquired images, a single preoperative 3D CT and a single-plane fluoroscopy at test time, making it readily applicable to various clinical settings.

Our work on vessel deformation prediction (section 4.4) served to validate the clinical viability of our approach, demonstrating its capability to augment fluoroscopic images. A key finding was that the network generalized well to testing data generated differently from the training data. While we identified a limitation in predicting displacement perpendicular to the image plane, this is mitigated by the fact that the resulting out-of-plane error is not visible in the augmented image.

Section 4.5 extended this work beyond breathing motion to handle various intervention-related deformations. This work best demonstrates the key strength of our approach, its adaptability to situations where deformation is unpredictable. Through extensive experiments, we validated our approach on experimental and synthetic data, and, in our sensitivity analysis, we measured the impact on performances of each component of our method. In our comparative study, we showed that our method performs competitively to the state-of-the-art in breathing motion recovery, and outperforms it in intervention-related deformation recovery.

Finally, in section 4.6, we integrated our work with V. Scarponi *et al.*'s guidewire navigation approach. Thanks to the vessels shape and position updates provided by our neural network, the success rate in moving anatomies was increased from 24% to 93%.

Our next steps, presented in chapters 5 and 6, will be to improve the realism of our predicted deformations, and more thoroughly validate our method on real fluoroscopic images.

Chapter 5

Physics-informed 2D-3D deformable registration

5.1 Introduction

The field of medical image registration first incorporated physics-based methods when Broit (Broit, 1981) introduced a linear elastic model for brain image registration in 1981. This pioneering work laid the foundation for subsequent biomechanical model-based registration methods. Since then, several biomechanical model-based registration methods have been published, a summary of which is proposed in Sec. 3.4. The advantage of these approaches is the incorporation of prior knowledge about the physical behavior of organs to deal with incomplete information. In our case, the lack of information manifests as a lack of contrast in fluoroscopic image, and in the ill-posedness of our problem, 2D-3D deformable registration.

In Sec. 5.2, our study on the use of the FEM to generate randomized, physically regularized deformations to train a 2D-3D registration network is reproduced. The study reports improved accuracy for the network trained on physically regularized deformations when tested on physically accurate deformations. Additionally, it assesses network performance on one sample of the IHUDeLiver10 dataset, which consists of pairs of CT scans of porcine subjects before and after intervention-related deformations. The key difference between the method used in this study and our baseline method, presented in Sec. 4.2, lies in the post-processing physical regularization step applied to generated displacement fields. This study was presented at the Data Curation and Augmentation in Enhancing Medical Imaging Applications (DCAMI) workshop of CVPR 2024 and was subsequently published in the workshop proceedings:

> François Lecomte *et al.* (June 2024). "Beyond Respiratory Models: A Physics-enhanced Synthetic Data Generation Method for 2D-

3D Deformable Registration". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2413–2421

In Sec. 5.3, we investigated the use of physics-based regularizers to enhance the realism of predicted deformations. First, we attempted to render the network prediction automatically differentiable with respect to spatial coordinates, to avoid using approximations in the regularization. Then, we evaluated using physics-based regularizers to train our network in conjunction with the loss on the displacement field.

5.2 Physics-based synthetic data generation for deformable registration

5.2.1 Introduction

We have previously presented several studies on domain-agnostic 2D-3D deformable registration (Chap. 4), where we generate randomized, diffeomorphic displacement fields (Sec. 4.2.2.1) to train our deformable 2D-3D registration network (Sec. 4.2.3). While this method ensures important geometric properties such as non-self-intersection ($J > 0$), it relies solely on geometrical constraints and does not guarantee physically realistic deformations. This limitation becomes particularly significant in our context of fluoroscopy-guided abdominal interventions, where real organs follow specific physical laws that restrict their possible deformations. The challenge is especially pronounced in regions with homogeneous CT intensity values, where the network must interpolate deformation fields from information gathered in high-contrast areas.

To address this, we propose incorporating physical constraints to regularize the displacement field generation process, a novel approach in fluoroscopy to CT deformable registration, where, traditionally, statistical deformation models derived from respiratory motion have been used (see Sec. 3.3.1). While maintaining most aspects of our previous methodology (Sec. 4.2), this study introduces an additional deformation post-processing step (detailed in Sec. 5.2.2) that ensures the physical plausibility of generated displacement fields within the organ of interest (in this case, the liver).

Furthermore, we implement several architectural modifications to our network compared to the version described in Sec. 4.2.3:

- In this work, only the reprojection loss presented in Sec. 4.2.3.2 was used.
- A Cosine Annealing learning rate scheduler (Loshchilov *et al.*, 2016), which decays the learning rate following a cosine schedule, was used.

- 8 encoder and decoder layers are used instead of 10
- The shape of the predicted and ground truth displacement fields is $(64, 32, 64)$ instead of $(128, 64, 128)$.
- The domain randomization data augmentation described in Sec. 4.2.4 was not used.

In all experiments, the network was trained for 30 epochs, with the learning rate set at 5.10^{-5}, which took approximately 2 hours on an Nvidia RTX 4090 GPU.

5.2.2 Physically-regularized displacement fields

While random DVFs generated with our baseline method are smooth and diffeomorphic, they may still incompletely represent the range of possible deformation during the fluoroscopy-guided intervention, which is the consequence of two main factors.

First, the parameters for the Gaussian kernels are sampled independently for each kernel, meaning that in any given DVF, there will be both large and small deformations. This is potentially different from real deformations, which may in some cases be small throughout the domain. Obtaining such a small deformation DVF from our randomized generation process is very unlikely, since it would require all realizations for α_k to produce small values. To remediate this, the generated DVF is multiplied by a scaling factor between -1 and 1, which ensures that samples with overall small displacements are better represented in the dataset.

Second, since the DVF generation process is stochastic, there is no guarantee that a body deforming under the influence of such DVF respects the conservation laws of physics. We therefore correct the DVF with a biomechanical model.

We are only interested in the deformation of the liver's internal structures (*e.g.* tumor, vessels), and therefore correct the DVF only inside the region occupied by the liver, hereafter denoted by $\mathbf{\Omega}$. To that end, in a preprocessing step, we first perform the liver segmentation and meshing to obtain a tetrahedral mesh representing the liver domain $\mathbf{\Omega}$. Then, the displacement inside the liver $\mathbf{U}$ is computed as the solution to the nonlinear elastostatic problem:

$$-2\nabla \cdot \frac{\partial \Psi}{\partial \mathbf{C}} = \mathbf{0}, \text{ in } \mathbf{\Omega} \tag{5.1}$$

where $\mathbf{C} = \mathbf{F}^T \mathbf{F}$ is the right Cauchy-Green deformation tensor, Ψ is the strain energy density function, and the gradient of deformation tensor $\mathbf{F}$ is related to the displacement field $\mathbf{U}$ via $\mathbf{F} = \nabla \mathbf{U} + \mathrm{I}$.

The liver is modeled as a hyperelastic Neohookean solid with strain energy density function:

$$\Psi = \frac{\lambda}{4}(J^2 - 1 - 2\ln(J)) + \frac{\mu}{2}(I_C - 3 - 2\ln(J)), \tag{5.2}$$

where $J = \det(\mathbf{F})$, $I_C = \mathrm{tr}(\mathbf{C})$ is the first invariant of the right Cauchy-Green deformation tensor, and μ and λ are the so-called *Lamé parameters*.

The Finite Element Method (FEM) was used to solve the elastostatic problem (5.1), with Dirichlet boundary conditions extracted from the DVFs prescribed at the liver boundary. The corrected DVF is then obtained by composing the physically accurate displacement solution $\mathbf{U}$ from (5.1) inside the liver, and the DVF outside the liver. Since all the liver's boundary is constrained, we used $\mu = 1$ and $\lambda = 0$ for all our biomechanical simulations.

5.2.3 Results

In order to validate our data generation approach, we evaluated the performances of a neural network trained on synthetic data, for two different registration contexts.

The first context was extracted from an open-source swine liver deformation dataset, IHUDeLiver10[1]. IHUDeLiver10 is composed of ten pairs of {*baseline; deformed*} Contrast Enhanced CT scans (CECT), experimentally acquired on ten different porcine subjects. For both images in each pair of CECT in the dataset, the portal vessel trees were segmented by an expert clinician, and serve to evaluate the registration accuracy. For each subject, the deformation of the anatomy was the result of a surgical procedure. Thus, this dataset contains realistic intervention-related deformations of the anatomy, which can be used to validate the effectiveness of our synthetic data generation approach. To transform the Contrast Enhanced CTs into regular, non-contrasted CTs, we used image inpainting (Barnes *et al.*, 2009) to remove as much of the contrast effect due to contrast agents as possible. The preoperative CT, *baseline* CT, was used to generate the training dataset, while the post-operative CT, *deformed* CT, was used to generate a test sample to evaluate the registration performance of the network. In this work, we only used one pair of experimentally acquired CT-Scans from the IHUDeLiver10 dataset (sample number 8). The deformation in the *deformed* CT of sample number 8 was induced by a surgical manipulation of the anatomy, reproducing deformations that may arise in a surgical intervention.

[1]IHUdeLiver10, along with data processing code, will be released at `https://doi.org/10.57745/EUBXGH`

The second context was generated synthetically, in order to test the accuracy of the method in a more controlled setting, less dependent on anatomical particularities that may influence the performance of the network. The *baseline* synthetic CT is composed of a cube of the same volume and at the same position as the liver of the first test case, surrounded by voxels with a constant intensity corresponding that of skin tissue. Inside the cube, the voxel intensities alternate along a checkerboard pattern, with tiles of side length 13.75 mm. The intensity values remain constant within each tile, but they gradually increase across tiles along the cube's main diagonal. For this case, the test samples were generated by setting constant Dirichlet boundary conditions on the left and right faces of the cube (while leaving the remaining faces stress-free), and solving the elastostatic problem (5.1) using the FEM with an hexahedral mesh of side length 10 mm. The displacement on the left face of the cube was set to 0, while the displacement on the right face of the cube was set to -40 mm, -20 mm, +20 m and +40 mm along the Left-Right (LR) axis, respectively. To generate the test samples, the cube was modeled as a hyperelastic Mooney-Rivlin solid, instead of the simpler NeoHookean solid used to generate the training data, in order to avoid bias regarding the choice of the hyperelastic model in the test data. The deformed mesh was then used to interpolate displacements on the CT image and produce the deformed CT scans. A DRR was then generated for each deformed CT scan, as described before.

For both registration contexts, the C-arm pose P was defined such that the projection is centered on the liver, and the viewing direction of the C-arm was aligned with the Antero-Posterior (AP) anatomical axis. In the following experiments, each dataset contains 18000 training samples and 2000 validation samples. Since the proposed use of the method is augmented anatomical visualization on 2D fluoroscopic images, all errors were measured on the 2D image plane after projection with the operator $\mathcal{P}$.

For the liver registration context, no point-to-point correspondences were available between the *baseline* and the *deformed* vessel trees, and we therefore chose the Earth mover's distance (EMD) metric to evaluate the registration accuracy. For the second registration context, the points of the cube mesh were used to measure registration accuracy directly. Since this test case is generated synthetically and the points are paired between the *baseline* and *deformed* images, we used the reprojection distance metric (RPD) which measures the euclidean distance after projection (using $\mathcal{P}$) on the image plane.

The Figs. 5.1 and 5.2 show the baseline and deformed DRRs for the liver and synthetic contexts, respectively.

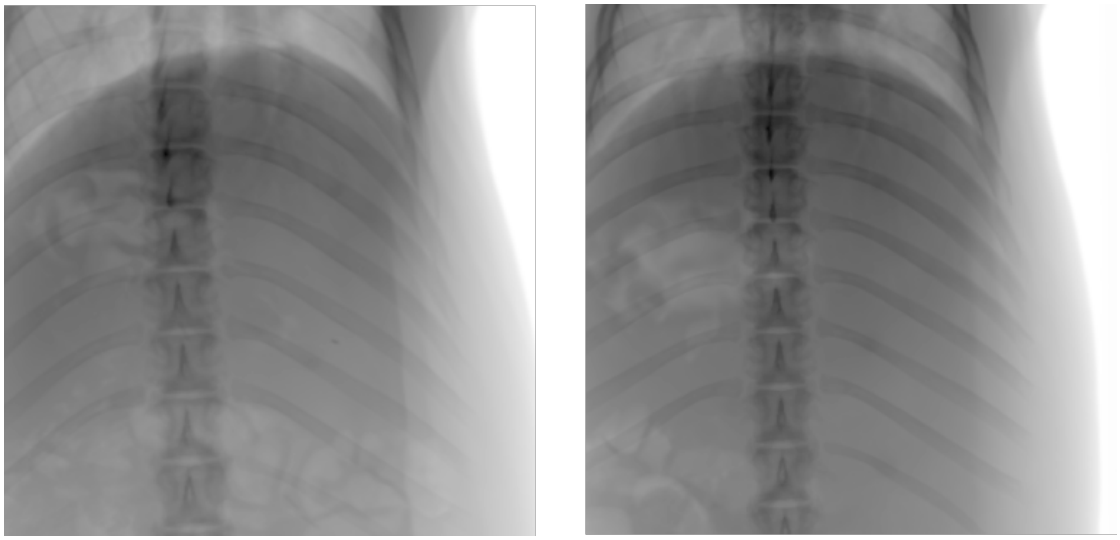


Figure 5.1: On the left, the DRR associated with the baseline CT and on the right the DRR associated with the deformed CT.

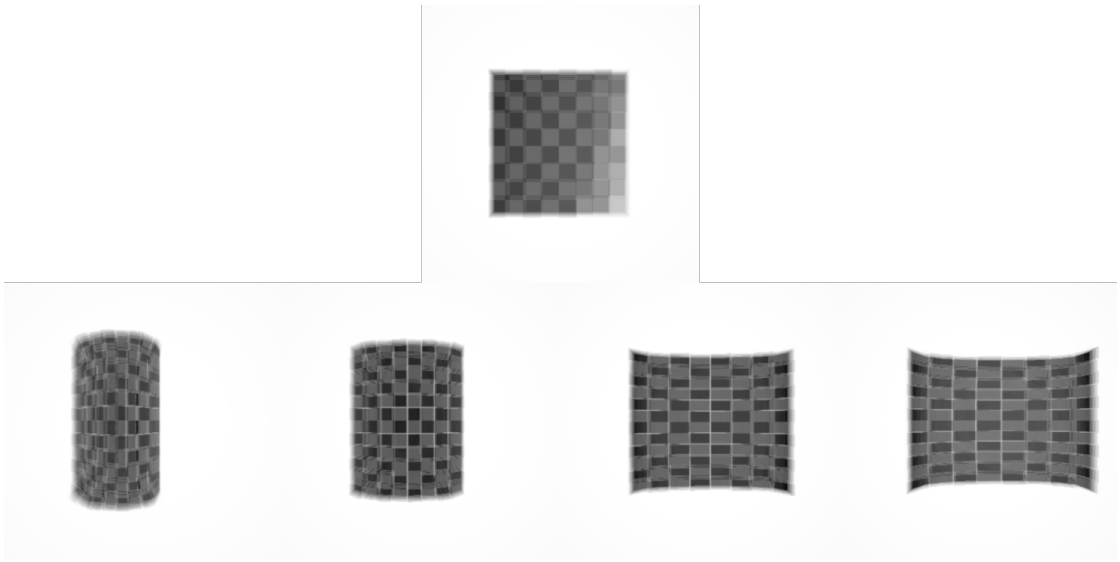


Figure 5.2: On top, the DRR associated with the baseline CT and on the bottom the DRRs associated with the deformed CTs, for displacements of -40 mm, -20 mm, +20 m and +40 mm (from left to right).

5.2.3.1 Registration accuracy

We evaluated the registration accuracy of the network trained using the synthetic data generation process described above.

The network was trained from scratch for each of the two test cases described above. For each test case, two training datasets were generated following Sec. 4.2.2.1, with and without the physical regularization step described in Sec. 5.2.2, in order to evaluate the effect of physics-based regularization.

For each test case, the registration accuracy of the network on the test sample(s) was measured every 3 training epoch. The tables 5.1 and 5.2 report the registration accuracy of the networks on the IHUDeLiver10 test sample and synthetic cube test samples respectively.

Epoch	Phy	Nophy
3	**4.0**	4.3
6	5.5	**3.6**
9	6.8	**4.7**
12	4.0	**3.9**
15	**4.2**	4.7
18	**3.4**	3.7
21	4.3	**3.3**
24	**2.8**	3.7
27	**3.7**	5.2
30	**3.6**	3.9

Table 5.1: Registration accuracy on the test sample every 3 epochs for networks trained with physically regularized (Phy) and not physically regularized (Nophy) data generation for the IHUDeLiver10 test case.

Beyond the 2D registration accuracy measurements presented in Tables 5.1 and 5.2, we extended our evaluation to examine performance in 3D space for both test cases, in Tables 5.3 and 5.4.

5.2.3.2 Ablation study

We performed two experiments on the IHUDeLiver10 test case to evaluate the impact of the data generation post-processing on the network performances. The architecture and training procedure of the network is the same for each experiment.

In the first experiment, three datasets were generated. The first dataset, termed "Base", was generated using the data generation process described above but without the post-processing described in Sec. 5.2.2. The second dataset, termed "Base + scale" was generated in the same way, but with the scaling post-processing and without the physical regularization. Finally, the third dataset, termed "Base + scale + phy" was generated using the full data generation process, with scaling and physical regularization, as described in Sec. 5.2.2.

Stretching amount (mm)	-40		-20		20		40	
Epoch	Phy	Nophy	Phy	Nophy	Phy	Nophy	Phy	Nophy
3	**17.79**	19.83	**5.81**	7.74	**5.10**	6.55	**18.61**	20.46
6	25.91	**24.77**	**4.96**	5.72	**4.87**	5.55	**19.51**	22.16
9	**14.13**	17.55	**3.50**	5.22	**3.82**	6.33	**17.49**	21.71
12	**15.02**	17.66	**4.20**	4.69	**4.32**	5.64	**17.52**	19.75
15	**16.67**	17.67	**4.26**	5.04	**4.29**	5.98	**17.84**	20.05
18	**17.18**	19.61	**5.12**	6.38	**4.39**	7.45	**18.36**	21.03
21	**18.50**	21.20	**4.28**	6.47	**4.34**	7.73	**19.28**	20.19
24	**18.31**	21.17	**4.48**	7.86	**4.32**	7.25	**18.86**	21.28
27	**18.81**	21.46	**4.42**	8.39	**4.68**	8.13	**19.87**	20.92
30	**18.83**	22.42	**4.40**	8.82	**5.25**	7.99	**20.13**	21.68

Table 5.2: Registration accuracy on test samples every 3 epochs for networks trained with physically regularized (Phy) and not physically regularized (Nophy) data generation for the synthetic cubes test cases.

	EMD 2D	CD 2D	EMD 3D	CD 3D
Before reg.	6.2	4.7	6.9	6.1
After reg.	2.8	2.0	4.7	4.1

Table 5.3: Earth Mover's distance (EMD) and Chamfer distance (CD) before and after registration on the sample 8 of the IHUDeLiver10 dataset, in mm.

The best registration performance for each dataset was 3.8 mm for the "Base" dataset, 3.3 mm for the "Base + scaling" dataset and 2.8 mm for the "Base + scaling + phy" dataset. In Fig. 5.3, the registration error of the network on the test sample is measured every 3 epochs.

In the second experiment, we used the "Base + scaling + phy" dataset to evaluate the effect of the number of training samples on the performances of the network. For each training run, only a portion of the dataset was used to train the network, from 0.1% to 100%. The results of this experiment are presented in Fig. 5.4.

Stretching amount (mm)	mean TRE	mean RPD
−40	11.35 (21.96)	14.13 (30.19)
−20	3.49 (10.77)	3.50 (14.91)
20	3.49 (10.53)	3.82 (14.70)
40	12.90 (20.91)	17.49 (29.28)

Table 5.4: Target Registration Error (TRE), in 3D, and Mean Reprojection distance (RPD), in 2D, on the synthetic cube dataset, in mm, with the error before registration in parentheses.

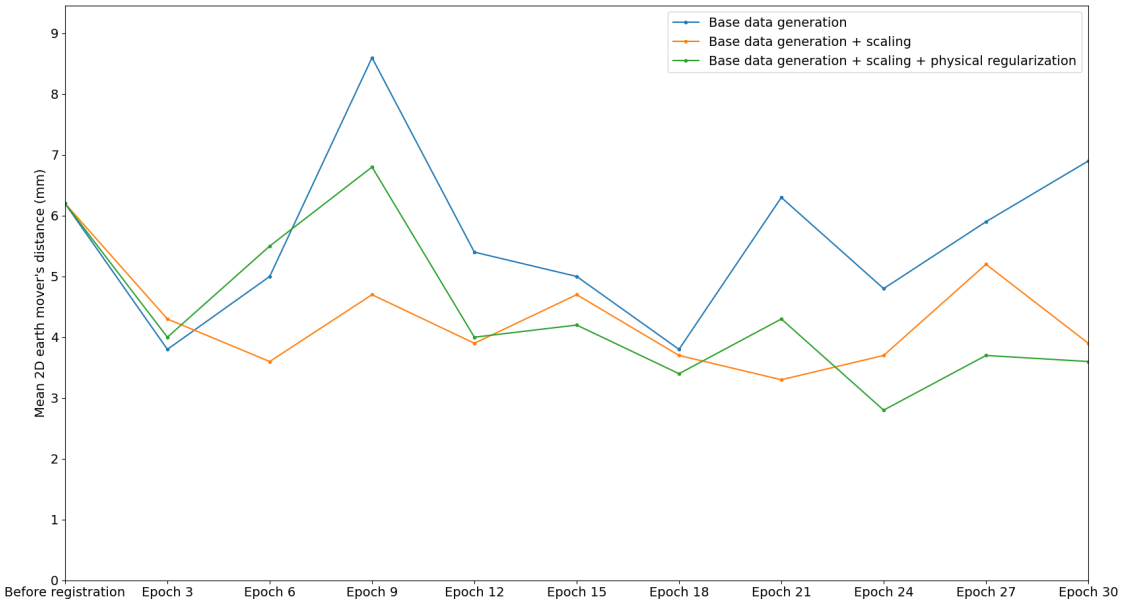


Figure 5.3: Each of the blue, orange and green curve shows the accuracy of the network every 3 epochs, for different data generation processes. In blue, the accuracy for the dataset generated following Sec. 4.2.2.1. In orange, the accuracy for the dataset generated with random scaling of the DVFs. In green, the accuracy for the dataset generated with the scaling and the physical regularization.

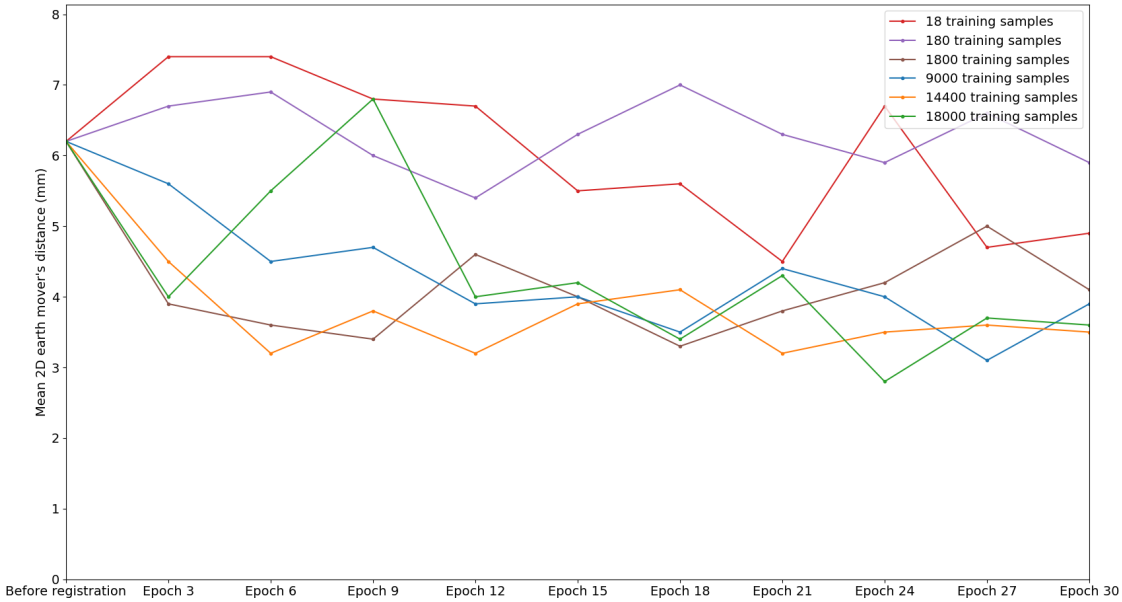


Figure 5.4: The red, purple, brown, blue, orange and green curves show the accuracy of the network every 3 epochs for dataset sizes of 18, 180, 1800, 14400 and 18000 respectively.

5.2.3.3 Qualitative results

The Fig. 5.5 shows the prediction of the best performing network, trained on physically regularized data for 24 epochs on the sample 8 from IHUDeLiver10 dataset. In most cases, the predicted vessel tree branches superpose well with the ground truth branches. Note that due to the experimental data acquisition process, the length of the branch is not the same between the baseline vessel tree and deformed vessel tree, making the numerical comparison between the predicted and ground truth vessel trees harder to evaluate.

The Fig. 5.6 shows the prediction of the best performing network, trained on physically regularized data for 9 epochs on the synthetic cube dataset. Displacements of $+20$ and -20 mm are well recovered, but larger displacements of -40 and 40 mm show that the network may be biased against large displacements. This may be due to the fact that larger displacements are less represented in the training dataset than smaller displacements.

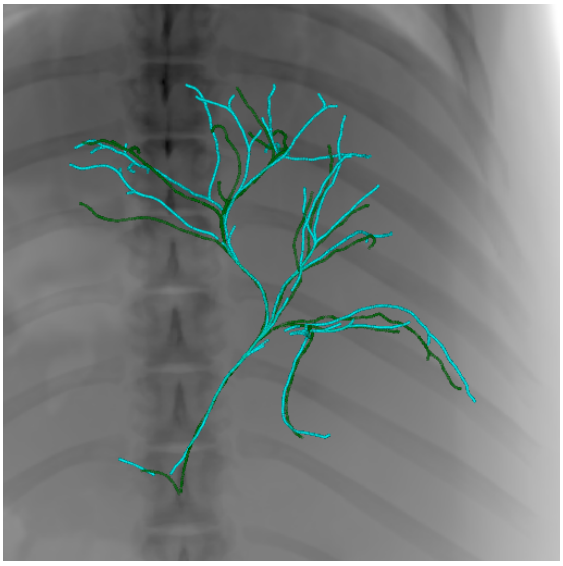


Figure 5.5: The vessel tree centerlines extracted from the deformed CT (in green) are overlaid on the DRR image generated from the deformed CT of the sample number 8 of the IHUDeLiver10 dataset. In blue, the vessel tree centerlines extracted from the baseline CT and deformed by the network prediction.

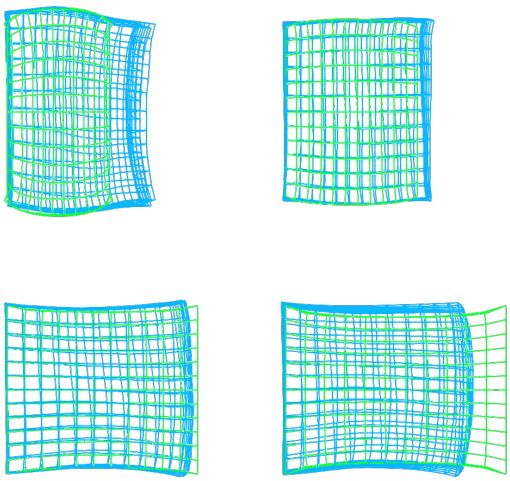


Figure 5.6: Top row: stretching of the cube of -40 mm (left) and -20 mm (right), with the ground truth mesh in green and the predicted mesh in blue. Bottom row: stretching of the cube of +40 mm (left) and +20 mm (right), with the ground truth mesh in green and the predicted mesh in blue.

5.3 Physics-based regularization

Building upon our physically regularized datasets from Section 5.2.3, we explored several approaches to further enhance the physical realism of predicted deformations. Our investigations focused on two main directions: improving the network architecture to enable efficient physics-based regularization and incorporating physics-based regularization during training.

An often-used regularizer, used to enforce smoothness, is the bending energy regularizer $\mathcal{L}_{bending} = ||\nabla \mathbf{F}||^2$, with $\mathbf{F}$ the gradient of the displacement field. As shown in previous studies (P. Alvarez and Cotin, 2024), this regularizer, while effective for smoothness, tends to penalize the magnitude of deformations and may lead the network to predict overly small deformations. To address this, Alvarez *et al.* propose a physically-motivated regularizer based on the hyperelastic strain energy density function of a NeoHookean material. This regularizer, $\mathcal{L}_{\nabla \cdot P}$ (Eqn. (5.4)), is based on the law of conservation of linear momentum (Eqn. (5.3), reproduced from (P. Alvarez and Cotin, 2024)).

$$\begin{aligned} \nabla \cdot \left(\frac{\partial \Psi}{\partial \mathbf{F}} \right) &= \nabla \cdot \mathbf{P} \\ &= \frac{\lambda}{2}(2J\mathbf{F}^{-T}\nabla J + (J^2 - 1)\mathbf{F}^{-T}) + \mu(\nabla \cdot \mathbf{F} - \nabla \cdot \mathbf{F}^{-T}) \\ &= \mathbf{0} \end{aligned} \tag{5.3}$$

With Ψ the NeoHookean strain energy density function and $\mathbf{F}$ the gradient of deformation, introduced in Eqn 5.2, $\mathbf{P}$ the linear momentum, λ and μ the Lamé parameters of the material, and $J = \det(\mathbf{F})$ the spatial jacobian.

$$\mathcal{L}_{\nabla \cdot P} = ||\nabla \cdot \mathbf{P}|| \tag{5.4}$$

For both of these regularizers, it is necessary to compute the spatial derivatives of the predicted deformation field. With our original architecture, spatial derivatives cannot be computed using automatic differentiation, but rather, must be approximated, for example using finite differences.

In Sec. 5.3.1, experiments are presented to render the network predictions automatically differentiable with respect to spatial coordinates, eliminating the need for derivative approximations in regularization. Then, in Section 5.3.2, we present our experimental results on incorporating physics-based regularization into the loss function.

5.3.1 Network architecture enhancement

In the first experiment, we developed a PointWiseDecoder variant that replaces the conventional decoder with a fully connected network. The fully connected network accepts a batch of spatial coordinates and outputs displacements at those coordinates, rendering predictions automatically differentiable with respect to spatial coordinates. To have the prediction depend on image features as well, we sample the feature map output by the transform module at the given spatial coordinates, and concatenate these features to the coordinates, before processing the concatenated tensor with the fully connected network. Another potential advantage of this variant is the possibility to train it only at specific points, such as the mesh points of an organ of interest, and simplify the network's task. Although inspired by Implicit Neural Representation approaches commonly used for solving physical equations (such as PINNs networks (Cai *et al.*, 2021)), or even for deformable registration by (Wolterink *et al.*, 2022), Wolterink *et al.*, the PointWiseDecoder variant obtained inferior performances to our baseline architecture, and we thus discontinued experiments on this variant. A possible explanation for the lower performance of this variant is the inability of a point-wise fully connected network to learn the mapping between coordinates and image features to a displacement field due to the lack of a long-range attention mechanism.

While Implicit Neural Representation approaches typically train networks to fit a single function (such as solving partial differential equations in PINNs or approximating displacement fields in (Wolterink *et al.*, 2022)), our scenario presents a distinct challenge. Our network must learn to predict mappings between pairs of functions, specifically from spatial coordinates and their encoded features to displacement vectors: $(\mathbf{x}, \mathit{Encoder}(\mathbf{x})) \rightarrow \boldsymbol{\varphi}(\mathbf{x})$. However, our use case is different, as the network needs to predict the mapping between two functions, $((\mathbf{x}, \mathit{Encoder}(\mathbf{x})) \rightarrow \boldsymbol{\varphi}(\mathbf{x}))$.

To reconcile our approach with the Implicit Neural Representation framework, we explored using a HyperNetwork architecture (see (Ha *et al.*, 2016)) to dynamically adjust the weights of the fully connected decoder, given the image features as input. In this way, for each sample, the decoder weights represent a single function, the displacement field to approximate. We implemented the HyperNetwork using fully connected layers that process feature maps from our 2D-3D transform module. Following (El Hadramy *et al.*, 2024), El Hadramy *et al.*, the network predicts a correction to the weights of the fully connected decoder. However, despite multiple attempts, this variant failed to achieve training convergence.

Given these challenges, we reverted to our original fully convolutional architecture for subsequent experiments, using finite differences to approximate spatial derivatives of predicted deformations.

5.3.2 Physics-based training regularization

We conducted extensive physical regularization experiments on both porcine and synthetic cube datasets (presented in Sec. 5.2.3), utilizing P. Alvarez's PyTorch implementation of $\mathcal{L}_{\nabla \cdot P}$. Our experiments involved:

- Testing multiple regularization weights to balance $\mathcal{L}\boldsymbol{\varphi}^{2D}$ and $\mathcal{L}_{\nabla \cdot P}$.
- Applying $\mathcal{L}_{\nabla \cdot P}$ selectively using a mask where $\mathcal{L}_{\nabla \cdot P} \simeq 0$ in the ground truth displacement field.
- Attempting pre-training with $\mathcal{L}\boldsymbol{\varphi}^{2D}$ alone to avoid early divergence of $\mathcal{L}_{\nabla \cdot P}$ when $J \to 0$.
- Implementation of a linear elasticity model of the liver for deformation regularization (Eqn. (5.5)).

Despite these various approaches, we were unable to achieve simultaneous convergence of both $\mathcal{L}_{\nabla \cdot P}$ and $\mathcal{L}\boldsymbol{\varphi}^{2D}$ during training.

$$\begin{aligned} \mathcal{L}^{lin}_{\nabla \cdot P} &= ||\nabla \cdot \mathbf{P}^{lin}|| \\ &= ||(\lambda + \mu)(\nabla \cdot \mathbf{F}^T) + \mu \nabla \cdot \mathbf{F}|| \end{aligned} \tag{5.5}$$

To check that $\mathcal{L}_{\nabla \cdot P}$ could indeed be used as a regularizer outside the Implicit Neural Representation framework, we designed a simplified registration experiment, where the goal is to perform register a binary 2D image, in the shape of a liver slice. The 2D slice underwent physically-regularized random deformations, generated as described in Sec. 5.2.2, using a 2D FEM implementation of the Neo-Hookean material. For deformation prediction, we designed a simplified variant of our original 2D-3D CNN architecture. The network maintained an encoder-decoder structure but omitted the 2D-3D transformation module, operating in 2D space only, with both encoder and decoder comprising eight convolutional layers each. We trained the network on a dataset composed of 18,000 training samples, as in previous experiments, and used a batch size of 32. The network was trained for 10 epochs with a fixed learning rate of 10^{-3}.

In our first experiment, we attempted direct network training using a combination of MSE loss on the displacement field and either the $\mathcal{L}_{\nabla \cdot P}$ hyperelasticity regularizer, or the $\mathcal{L}^{lin}_{\nabla \cdot P}$ linear elasticity regularizer, testing various regularizer weights. We scaled the network prediction by $10^{-}4$ to ensure small initial predictions after weight initialization, preventing $\mathcal{L}_{\nabla \cdot P}$ divergence, and applied a mask to compute predictions only in the organ region. This approach proved unsuccessful, and Fig. 5.7 illustrates why: $\nabla \cdot P$ values grow rapidly for non-physical

deformations, with displacement predictions at scale 10^{-3} resulting in $\nabla \cdot P$ values at scale 10^{1}. Without prediction scaling, some $\nabla \cdot P$ values become undefined due to negative J values for large, random deformations, leading to NaN values in the loss function. Moreover, since $\mathcal{L}_{\nabla \cdot P}$ grows non-linearly with displacement amplitude, using a small regularization weight proved insufficient for joint minimization of both $\mathcal{L}_{\nabla \cdot P}$ and MSE loss during training.

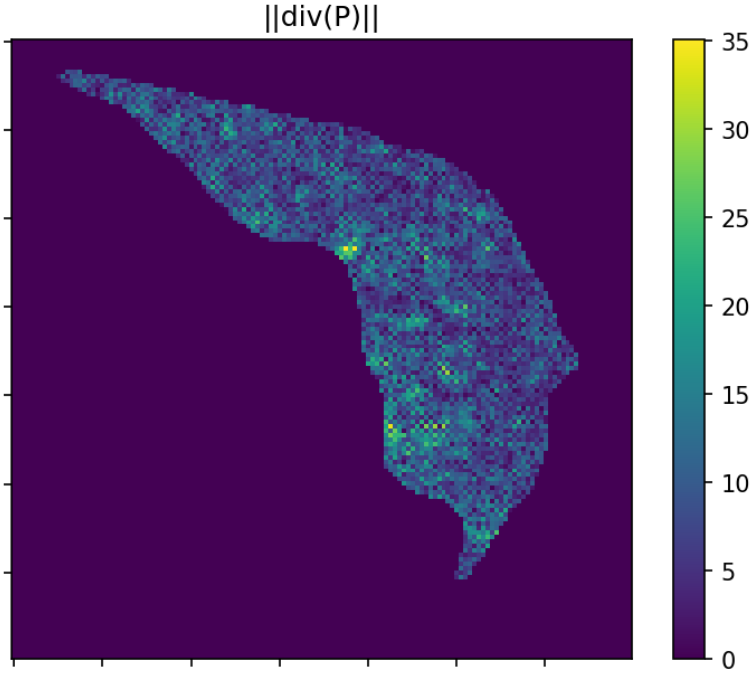


Figure 5.7: Values of $\mathcal{L}_{\nabla \cdot P}$ of the predicted displacement field at weight initialization.

In our second experiment, to address the issue of large $\mathcal{L}_{\nabla \cdot P}$ values disrupting training, we initially trained the network using only MSE loss on the displacement field for 10 epochs. We then continued training with either $\mathcal{L}_{\nabla \cdot P}$ or $\mathcal{L}^{lin}_{\nabla \cdot P}$. Training with $\mathcal{L}_{\nabla \cdot P}$ failed to converge regardless of regularization weight. However, training with $\mathcal{L}^{lin}_{\nabla \cdot P}$ successfully converged, with $\mathcal{L}^{lin}_{\nabla \cdot P}$ minimized without increasing MSE loss, as shown in Fig. 5.8. This experiment used a regularization weight of $1 \cdot 10^{-5}$, to account for scale differences between the MSE loss and $\mathcal{L}^{lin}_{\nabla \cdot P}$, and a reduced learning rate from $1 \cdot 10^{-3}$ to $1 \cdot 10^{-4}$.

In our third experiment, we retrained the network trained with $\mathcal{L}^{lin}_{\nabla \cdot P}$ using $\mathcal{L}_{\nabla \cdot P}$ with a regularization weight of $1 \cdot 10^{-5}$ and a further reduced learning rate of $2 \cdot 10^{-5}$. This approach successfully minimized $\mathcal{L}_{\nabla \cdot P}$ without increasing MSE loss, as demonstrated in Fig. 5.9. This experiment demonstrates that $\mathcal{L}_{\nabla \cdot P}$ can

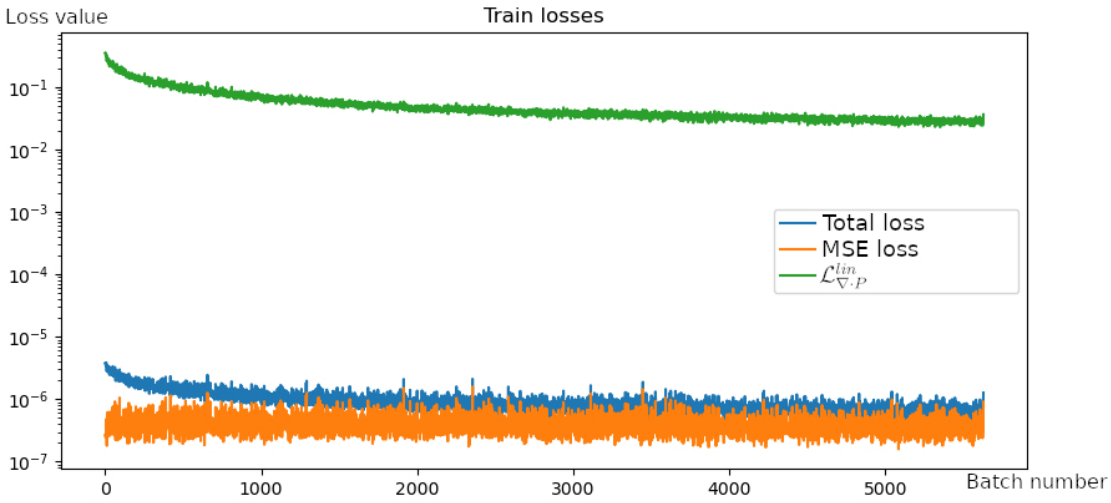


Figure 5.8: Values of the MSE loss and $\mathcal{L}^{lin}_{\nabla.P}$ during retraining.

indeed be used to regularize training, although requiring two pre-training steps, first using an MSE and then a linear elasticity regularization.

Finally, we trained the network from scratch using only MSE loss, replicating the training process of the $\mathcal{L}_{\nabla.P}$-trained network to enable meaningful comparison. Fig. 5.10 shows the measured $\mathcal{L}_{\nabla.P}$ values for network predictions on the validation dataset in each case. Additionally, Fig. 5.11 provides a qualitative comparison of network predictions with and without regularization.

5.4 Discussion and conclusion

Physics-based synthetic data generation

In Sec. 5.2.3.1, the registration accuracy of the network trained on synthetic deformations was evaluated with and without physical regularization. On the porcine test case from the IHUDeLiver10 dataset, the best registration performance is 2.8 mm, obtained at epoch 24 for the network trained on physically regularized data. However, the accuracy of the network does not improve monotonically during training, suggesting that early stopping may be necessary to obtain the best registration performances on the test set. Additionally, while the physical regularization generally improves performances, it is not true for all epochs. Finally, this experiment would need to be repeated on the full IHUDeLiver10 dataset to better appreciate the registration performances of the method.

On the synthetic cube test case, the difference in accuracy between the networks trained on physically regularized and not physically regularized data is more clear, with the physically regularized method performing almost always better. This

may be related to the relative lack of contrast of the synthetic dataset, with each tile of the checkerboard pattern being of constant intensity. Without contrast, the deformation inside the tile can only be inferred from the deformation of the tile edges. With the physical regularization, the network might be able to learn to better interpolate the displacement inside the tile from the displacement at the edges of the tile. Again, while there is no monotonic convergence, the best results are still obtained with the physically regularized data.

In Sec. 5.2.3.2, the first ablation study experiment shows the importance of adjusting the synthetic training data distribution to better match the testing data distribution. Despite its simplicity, removing the "scaling" transformation resulted in very poor registration performances, with the network failing to converge. On the other hand, the addition of the biomechanical regularization, which induces a non-negligible additional computational cost, improved the registration performance by a modest amount, and only at some training epochs. However, there are other aspects to take into consideration for physically regularized registration, namely choosing the right biomechanical model with the right parameters, and choosing physically plausible boundary conditions. In our cases, while the random DVF is smooth and diffeomorphic, it does not respect the conservation laws of physics. Due to this, the surface of the organ may be subject to physically implausible deformations, giving rise to unrealistically high strain energy inside the organ. However, despite these limitations, the best performance is attained by the network trained on the physically regularized dataset, with a clinically relevant accuracy of 2.8 mm (from 6.2 mm before registration).

The second ablation study experiment sheds some light on the number of training samples necessary to learn the 2D-3D registration task. We found that networks trained with 18 or 180 samples consistently produced errors above 4 mm. Performance significantly improved with larger training sets, with optimal results achieved using 18,000 samples. Notably, networks trained on 1,800 to 18,000 samples showed comparable performance levels, suggesting that increasing the dataset size beyond 18,000 would offer limited performance gains. This indicates that factors such as network architecture, training strategy, and data generation methodology may be more crucial for improving performance than increasing dataset size.

Finally, 3D errors measurements in Tables 5.3 and 5.4 reveal that while registration improves the overall alignment, the 3D accuracy falls below the 2D performance levels. This limitation stems from the 2D image-based training approach: since out-of-plane motion remains invisible in the 2D training images, it isn't captured by the loss function, leaving the network insensitive to deformations in this direction. This limitation stems from the inherent loss of information in the 3D to 2D projection operation, fundamentally limiting our registration approach. Since out-of-plane motion information is lost in this projection process, the network's

loss function cannot account for deformations in this direction. While this limitation is acceptable for our target application of interventional augmented 2D visualization, biomechanical regularization could be employed in the loss function as well to better estimate out of plane motion.

Physics-based regularization

Our investigation into making network predictions spatially differentiable yielded inconclusive results (Sec. 5.3.1). While this property would theoretically facilitate physical regularizer computation and potentially eliminate training instabilities, it represents a novel approach not yet thoroughly explored in literature. Particularly, a specifically designed network architecture able to capture long-range spatial relationships in the displacement field could be employed to tackle this problem.

In Sec. 5.3.2, we implemented and used physical regularizers to train our 2D-3D deformable registration network, without success. The challenges encountered suggest difficulties in combining physical regularization with the complexities of 2D-3D deformable registration. We thus redirected our investigation to a simpler 2D registration problem.

Initial attempts to train from scratch with physical regularization proved unsuccessful, highlighting the inherent difficulties in simultaneously optimizing registration accuracy and physical plausibility. However, we achieved encouraging results through an incremental training approach. This strategy involved first training a network to predict reasonably smooth displacement fields, then progressively incorporating physical constraints. We first retrained with the more stable linear elasticity regularizer ($\mathcal{L}^{lin}_{\nabla \cdot P}$), followed by the hyperelastic regularizer ($\mathcal{L}_{\nabla \cdot P}$). This sequential approach successfully produced physically plausible and smooth predicted deformations, particularly valuable in regions lacking contrast information, as demonstrated in our visual comparison in Fig. 5.11. Our quantitative analysis in Fig. 5.10 revealed that while training with $\mathcal{L}^{lin}_{\nabla \cdot P}$ alone could achieve generally physically plausible predictions, the more accurate $\mathcal{L}_{\nabla \cdot P}$ regularizer yielded superior results

The failure of direct integration of physical regularization in our 2D-3D registration network, combined with the eventual success in the simplified 2D case, indicates that the complexity of simultaneous 2D-3D registration and physical constraint learning may exceed our current architecture and training procedure capabilities.

This work demonstrates that physical regularization can improve network predictions, though our investigation reveals it's not a straightforward process. Our preliminary results, especially in 2D experiments, highlight the potential benefits of this approach in contexts with limited information, such as low-contrast images or binary 2D representations. Future research could focus on incorporating phys-

ical constraints directly into the network architecture, potentially guaranteeing the physical realism of predicted deformations even before training. This architectural approach, as proposed in emerging work on Physics-Augmented Neural Networks (Linden *et al.*, 2023), might offer a more robust solution than post-hoc regularization. Additionally, developing more stable training procedures and investigating methods to extend successful 2D approaches to full 2D-3D registration problems remain important areas for investigation.

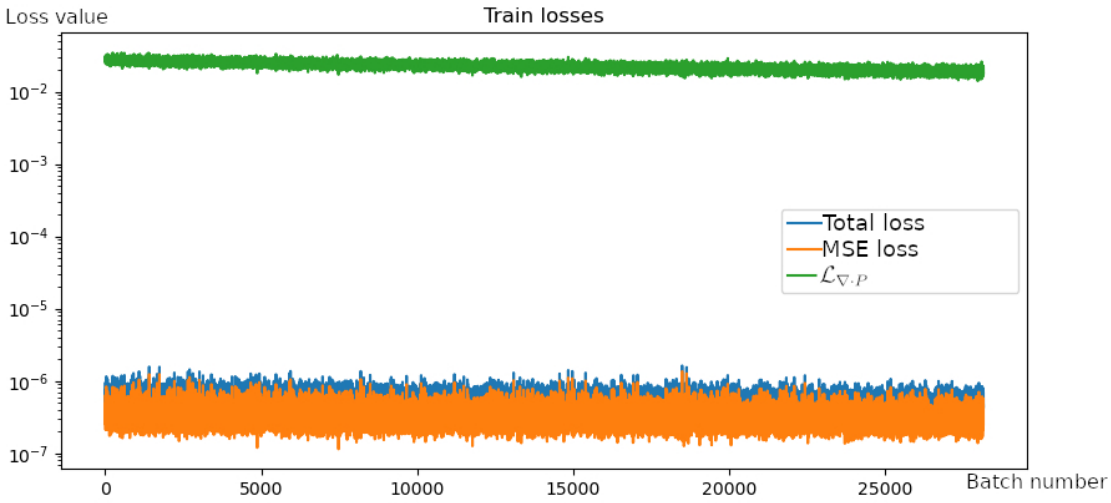


Figure 5.9: Values of the MSE loss and $\mathcal{L}_{\nabla \cdot P}$ during retraining.

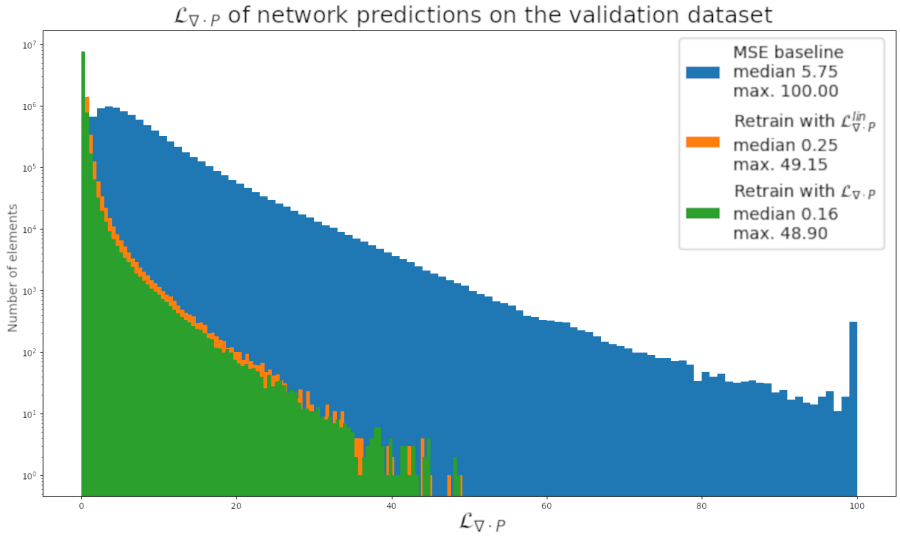


Figure 5.10: Histogram of $\mathcal{L}_{\nabla \cdot P}$ of network predictions on the validation dataset for different training setups.

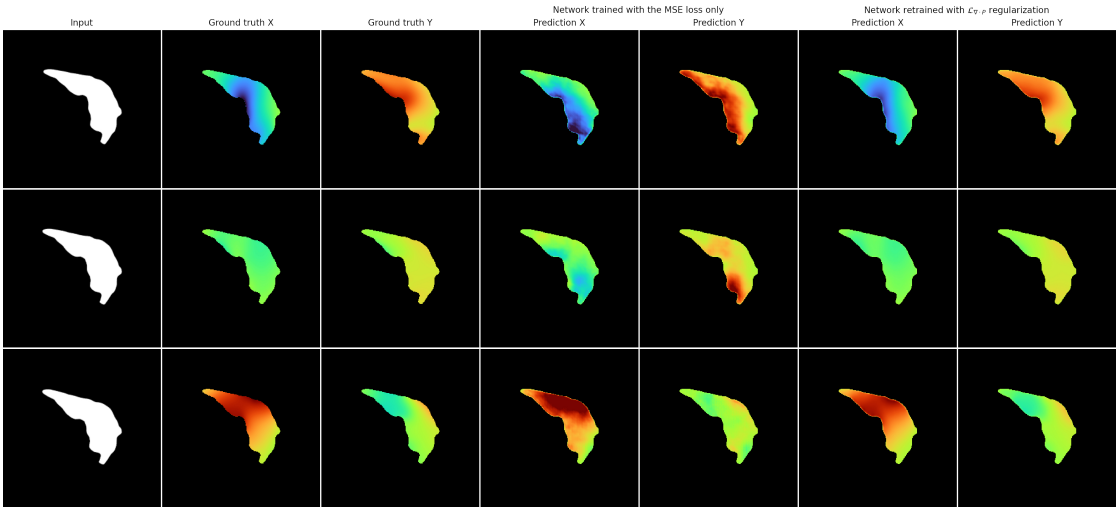


Figure 5.11: Qualitative comparison between network predictions with and without retraining with $\mathcal{L}_{\nabla \cdot P}$.

Chapter 6

2D-3D deformable registration in the real world

6.1 Introduction

Fluoroscopy-guided interventions currently rely on CBCT-fluoroscopy imaging as the gold standard for accurate anatomical localization. CBCT devices, as discussed in section 1.2.3, are employed to capture multiple 3D scans throughout the intervention. In fluoroscopy mode, these CBCT scans can be rigidly registered and superimposed onto real-time fluoroscopic images to provide enhanced guidance. When deformations occur, this enhanced guidance becomes outdated and another CBCT acquisition is required to update it, causing significant radiation exposure over the course of the intervention. Thus, a transition to fluoroscopy-only augmented guidance would offer several significant advantages over the current CBCT-fluoroscopy approach, notably, reduced radiation exposure, continuous real-time updates of augmented visualizations, and no dependency on costly CBCT equipment.

However, transitioning to fluoroscopy-only guidance requires an accurate, real-time 2D-3D deformable registration algorithm. Before our 2D-3D deformable registration approach could be used for fluoroscopy-only augmented guidance, two primary challenges must be overcome. The first involves determining the pose of the fluoroscopic imaging device relative to the preoperative CT anatomy. To find this pose, 2D-3D rigid registration approaches have been developed over the years, as presented in 3.2. In our experiments, we used one such approach, DiffPose, developed by (Gopalakrishnan, Dey, *et al.*, 2024), Gopalakrishnan, Dey, *et al.*, to estimate the pose of experimentally acquired fluoroscopic images.

The second challenge lies in bridging the domain gap between synthetic Digitally Reconstructed Radiographs (DRRs) and actual fluoroscopic images. This

gap manifests in multiple image characteristics, including:

- Variations in contrast and exposure
- Different noise patterns and resolution
- Anatomical discrepancies between preoperative and intraoperative states due to:
  - Bowel movements
  - Changes in patient positioning
  - Partial lung atelectasis
  - etc...

To enhance our network's robustness to these variations, we used domain randomization to augment the input images during training (see Sec. 4.2.4).

To validate our approach, we conducted experiments on two experimentally acquired datasets. Given inherent limitations about ground truth deformations in both datasets, our evaluation remains semi-quantitative. The common elements of our experimental setup are detailed in Section 6.1.

Section 6.3 presents our evaluation using a porcine dataset featuring radiopaque markers, where respiratory motion induces anatomical deformation in the fluoroscopic images. Since this dataset does not contain ground truth CT scans associated with deformed fluoroscopic images, we are only able to evaluate the 2D accuracy of our method, rendering the evaluation semi-quantitative in nature.

Section 6.4 describes our experiments with clinical data obtained from minimally invasive lung interventions at Rennes CHU. This dataset comprises a preoperative CT scan, two intraoperative CBCT volumes, and their associated fluoroscopic projections. While the presence of a nodule in both CT and CBCT volumes provides a reference point for registration evaluation, two significant challenges affect our assessment. Firstly, the ground truth poses of fluoroscopic projections relative to the preoperative CT volume are unavailable. Secondly, the preoperative and intraoperative volumes exist in different reference frames, necessitating an additional, error-prone, 3D-3D rigid registration step to enable the computation of errors in 3D. These limitations collectively constrain our ability to perform fully quantitative validation, leading us to characterize our experimental results as semi-quantitative.

6.2 Experimental setup

For each dataset, we first estimated the pose of the real fluoroscopic images with respect to the preoperative CT, as detailed in Sec. 6.2.1. Then, using the estimated

poses, we generated synthetic datasets from the preoperative CT image, following the process described in Sec. 4.2.2 and 6.4.1 for the porcine and interventional datasets, respectively. The accuracy of our method was evaluated across several variations, with results presented in Sec. 6.3.1 and 6.4.2.

6.2.1 Pose estimation

Initial pose estimation was based on anatomical landmarks, specifically the spine and ribs, visible in the fluoroscopic images. For the porcine experiment (Sec. 6.3), the central position of the spine and horizontal orientation of the ribs indicated an anteroposterior projection axis. In the clinical experiment, presented in Sec. 6.4, we selected the projection halfway in the fluoroscopic sequence, acquired along a semicircular trajectory around the patient, as the reference projection. In this projection, the spine's left-sided position and shortened, curved rib segments suggested a Left-Right projection axis.

We refined these initial pose estimates through a three-step process:

1. Manual refinement using the DiffDRR renderer (Gopalakrishnan and Golland, 2022), iteratively adjusting translation and rotation parameters until the generated DRR visually matched the reference fluoroscopic image as closely as possible.

2. Initial automated refinement using DiffPose (Gopalakrishnan, Dey, *et al.*, 2024) For this step, we trained a ResNet18 convolutional neural network on a synthetic dataset generated with randomized poses near the initial estimate. This step failed, and we thus used the pose found manually to initialize the next step.

3. Final refinement using DiffPose's iterative optimization module. This step employed the DiffDRR renderer to automatically optimize pose parameters by minimizing a structural similarity index measure (SSIM)-based loss.

This three-step process achieved high accuracy for the porcine dataset, with generated DRRs closely matching the fluoroscopic images. For the clinical dataset, this process failed and we used the pose found manually. The final estimated poses are shown in Fig. 6.1. The resulting poses served as baselines for data generation, as detailed in Sec. 4.2.2 and 6.4.1.

6.3 Experiments on a porcine model

For this experiment, an experimentally acquired porcine dataset comprising a preoperative CT scan with 12 implanted radio-opaque markers and a sequence of 74

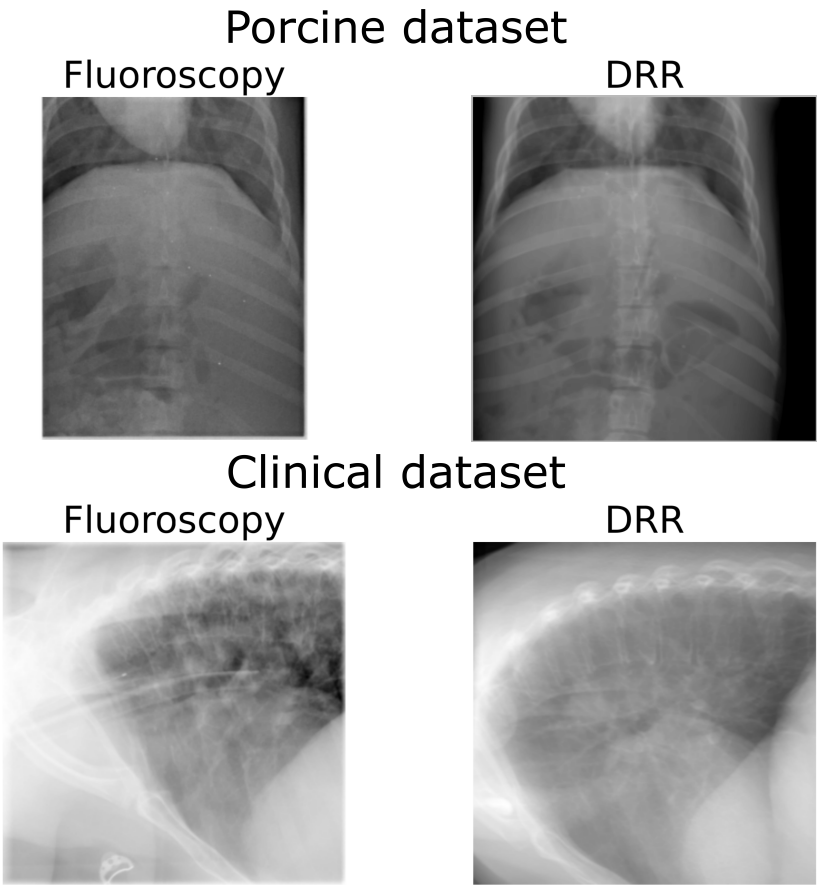


Figure 6.1: DRRs generated with poses estimated from fluoroscopic images for the porcine (top) and clinical (bottom) datasets.

fluoroscopic images captured during breathing was used. This dataset contains a pre-operative CT scan, with 12 implanted radio-opaque markers, Following pose estimation (Sec. 6.2.1), we evaluated several variations of our method on this dataset (Sec. 6.3.1).

6.3.1 Results and discussion

Since the markers are present in both the CT and fluoroscopic images, they can be used to measure the registration accuracy of our method in 2D. This constitutes a semi-quantitative evaluation, because, unfortunately, we do not have access to the 3D positions of the markers during breathing. A potential solution would be to acquire and temporally synchronize a 4D CT sequence with the fluoroscopic image sequence.

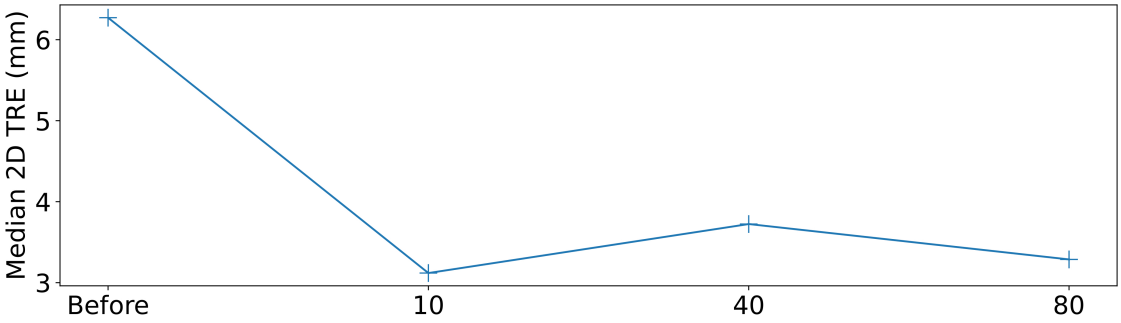


Figure 6.2: Network accuracy for varying numbers of total training epochs. 'Before' refers to the initial error between markers in the preoperative data and ground truth markers.

To automatically extract the marker's positions in fluoroscopic images, we used the Segment Anything (SAM) pretrained model (Kirillov *et al.*, 2023). Given the manual annotation of marker positions in one image, SAM automatically segmented the markers in all other images in the sequence. Markers positions were defined as segmentation centroids, and verified through visual inspection. To ensure accuracy, we plotted the temporal evolution of markers' vertical and horizontal positions and eliminated frames showing abrupt position changes. This methodology successfully identified the positions of eight liver-implanted markers in 73 out of 74 frames.

Our method was evaluated through four experiments on this dataset. In the first experiment, we investigated the relationship between network performance and training duration. As introduced in Sec. 4.2.3, we used the OneCycleLR policy to vary the learning rate during training. With this policy, over the course of training, the learning rate follows a cosine curve, starting from $\eta = \eta_{min} = \frac{\eta_{max}}{10^4}$, increasing to $\eta = \eta_{max}$ and then decreasing back to η_{min}. When the number of epochs varies, the period of the cosine changes in consequence such that η always follows the same curve during training. The results of this experiment are summarized in Fig. 6.2, where the 2D error on the landmarks is measured for different numbers of training epochs. Optimal performance was achieved with 10 epochs, though using a non-optimal batch size of 16, as we will see in the last experiment.

The second experiment addressed an implementation issue discovered during feature map visualization. We identified that the 2D to 3D feature reshape operation in the backprojection module wasn't correctly preserving the 3D mask (see Fig. 6.3). This is due to the fact the depth dimension can be split into a depth and a feature dimension in two different ways, either with the feature dimension first or last. Depending on this choice, the output of the reshape operation is largely

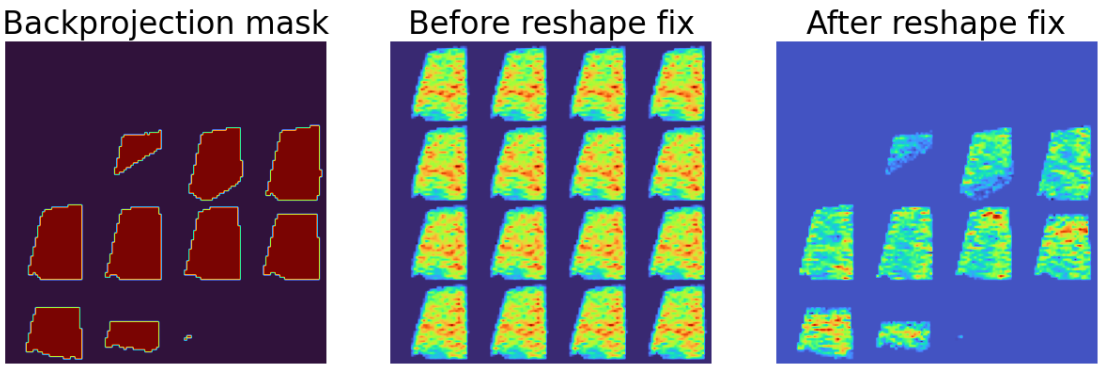


Figure 6.3: From left to right, the 3D mask used in backprojection, the 3D features after backprojection, masking, and reshaping (see Sec. 4.2.3.1), and the 3D features after backprojection, masking, and reshaping with the fix implemented. This experiment was performed with four decoder layers to increase the resolution of feature maps and make visualization easier.

different. The fix consists in splitting the depth dimension into a depth dimension first and a feature dimension last rather than feature first and depth last.

Consequently, we implemented a fix to modify the reshape operation in the backprojection module. We also hypothesized that masking the feature maps after backprojection could potentially remove important information, and we added an option to disable the masking operation to test this hypothesis. Since this effect is more prevalent at higher resolutions, this comparison between architectures was performed with only four decoder layers, to increase the initial 3D feature maps resolution to $(32, 16, 32)$ instead of $(4, 2, 4)$ with ten decoder layers. We tested three configurations: the original implementation (NoFix), a corrected version (Fix), and a version with disabled post-backprojection masking (FixNoMask), using a batch size of 16 each time. This experiment supported our hypothesis, as the median 2D error improved from 4.5 mm (NoFix) to 3.4 mm (Fix) and 3.4 mm (FixNoMask). However, with ten decoder layers, FixNoMask performed worse (3.8 mm) compared to NoFix (3.1 mm), suggesting that a change of architecture of the network could improve the performances of the FixNoMask variant, to keep both a high spatial resolution after backprojection and a sufficient number of layers in the decoder. To achieve this, in the decoder, the features could be spatially upsampled by a factor of two only once every four layers instead of every two layers, effectively increasing the spatial resolution of backprojected features for a fixed output resolution.

In a third experiment, we tried to address the consistent registration failure for one marker in all experiments. In DRRs generated from the preoperative CT, this marker is superimposed with partially full intestines, while in the test fluoroscopic images, the intestines show a different appearance, leading to a change in image

Figure 6.4: From left to right, a DRR generated from the preoperative CT, a fluoroscopic image from the test set, the same fluoroscopic images with superimposed preoperative (blue) and ground truth (green) marker positions. The red circle highlights the region around the marker for which registration fails, which shows a difference in image appearance possibly unrelated to a deformation.

appearance at the marker position. Furthermore, the marker position in the test images never coincides with the preoperative position, suggesting that a shift has occurred, either of the marker relative to the liver (migration) or of the liver in this region. A comparison between a preoperative DRR and a testing fluoroscopic image is presented in Fig. 6.4, with the aforementioned marker highlighted. To check if the change in appearance was responsible for the systematic error on the position of this marker, we generated a new training dataset with randomized intensity perturbations to simulate changes in image appearance unrelated to deformations. To create these perturbations, for each sample, 0 to 10 cubic regions of uniform low or high intensity and randomized radii were created in the CT data before deformation and DRR rendering. Intensity values in the liver, the organ of interest here, were preserved. An example of a DRR containing perturbations is presented in Fig. 6.5. The goal of this training dataset was to make the network more robust to changes in image intensities unrelated to deformations, such as the stomach or intestine contents changing due to digestion. This hypothesis was not verified after the training process, which led to a median 2D error of 4.2 mm, suggesting that the registration failure might stem from actual marker displacement rather than appearance changes. In the future, experiments on the failure modes of our method would be necessary to better define the clinical use cases of our method.

Finally, in a fourth experiment, we studied the impact of batch size on performance. To train our network, we have access to either an Nvidia RTX 4090 (24 GB memory) supporting batch size 16, or an Nvidia GeForce GTX 1080 Ti (11 GB memory) limited to batch size 7. On the first GPU, we were able to use

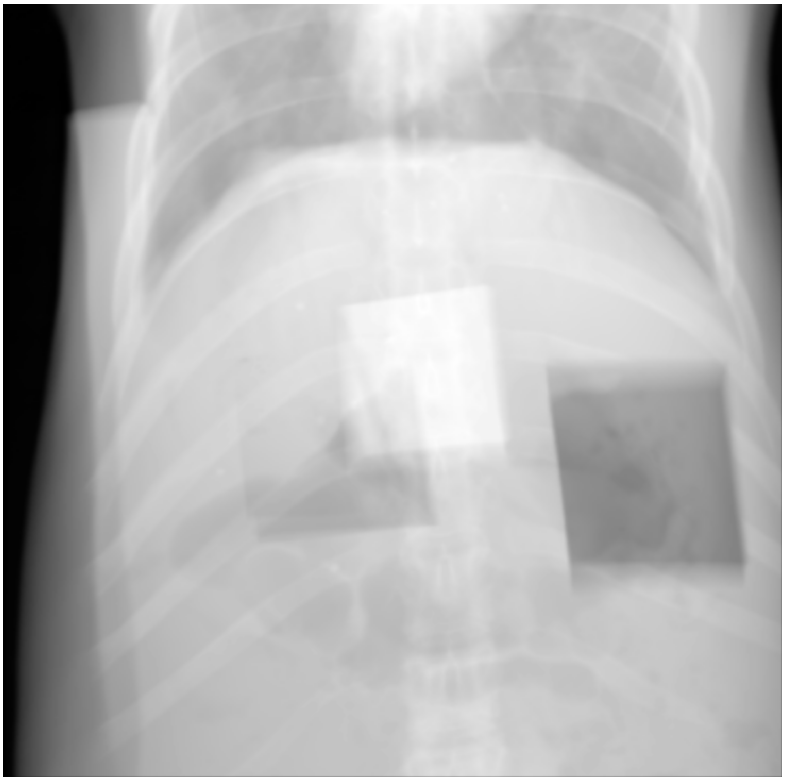


Figure 6.5: A DRR generated with randomized perturbations to simulate image intensity changes unrelated to deformations. Perturbations are cubic regions of varying sizes with uniform intensities and are meant to simulate intensity changes due to anatomical processes such as digestion.

a maximal batch size of 16, while on the second GPU, the batch size had to be reduced to 7. Previous research shows conflicting conclusions about batch size impact on CNN performance, with some works concluding that accuracy diminishes with increasing batch sizes (F. He *et al.*, 2019; Kandel *et al.*, 2020), and others drawing the opposite conclusion (Radiuk, 2017). This experiment aims at determining the optimal batch size for our application, in terms of performance of the trained network on testing data. The results of this experiment are reported in Fig. 6.6, and show that, optimal results are obtained with a batch size of 7, with an improved median error of 2.4 mm, from 3.1 mm with a batch size of 16.

6.4 Experiments on clinical data

In this experiment, we utilized a clinically acquired dataset from a patient undergoing lung nodule resection. This dataset includes a pre-operative CT scan,

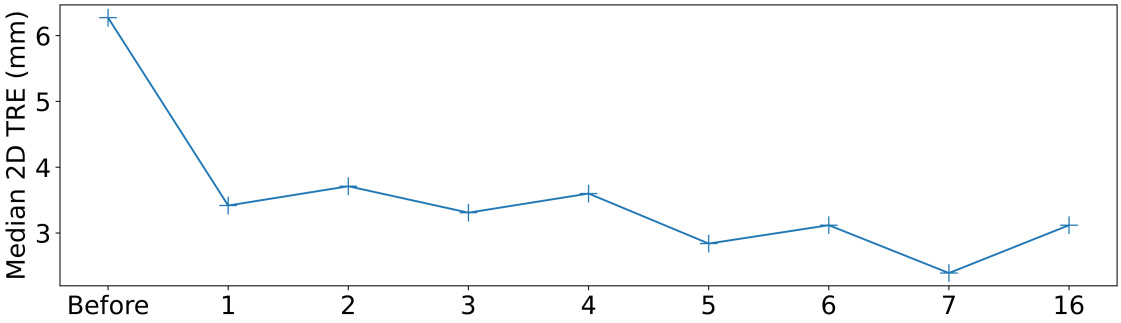


Figure 6.6: Network accuracy for different batch sizes.

showing a nodule in the upper part of the right lung, and an intraoperative CBCT scan with associated fluoroscopic images, acquired at the start of the intervention. The patient's position differs between the scans: in the CBCT scan, the patient is in the lateral decubitus position (lying on their left side), while in the preoperative CT scan, the patient is in the supine position (lying on their back). This patient pose change induces a deformation of the anatomy, which has been the subject of prior 3D-3D registration approaches by Alvarez *et al.* (P. Alvarez, Chabanas, *et al.*, 2022; P. Alvarez, Rouzé, *et al.*, 2021; P. A. Alvarez, 2020) and Boussot *et al.* (Boussot *et al.*, 2023; Boussot *et al.*, 2022). To test the ability of our method to recover this deformation from a single fluoroscopic image, we used the fluoroscopic images associated with the CBCT scan as test data. These images were acquired sequentially while the CBCT device rotated approximately 130 deg around the patient. To recover the pose of these images, we selected the image from the middle of the sequence, which was acquired at an angle perpendicular to the operating table. We then performed the pose estimation process described in Sec. 6.2.1 on this image, which yielded suboptimal results (see Fig. 6.1), possibly due to differences in image appearance between DRR images and fluoroscopic images. To expand the testing dataset and assess our method's robustness to different poses, we also used fluoroscopic images within a 30 deg range around the recovered pose (along the rotation axis), and computed the corresponding poses. The following sections (Sec. 6.4.1 and Sec. 6.4.2) will provide details on the training data generation process and present the experimental results.

6.4.1 Data generation

As previously mentioned, each test fluoroscopic image in this dataset is acquired with a different pose. Additionally, the pose estimation process did not yield optimal results, making it necessary for the network to be robust against pose variations. To achieve this, training DRRs were generated with poses varying around the previously found poses. Two datasets were generated: P020_p30 and

P020_small.

The P020_p30 dataset was generated with large variations around the poses generated in the 30 deg range around the base pose, with translation amplitudes following a half-normal distribution (standard deviation $\sigma = 100$ mm, amplitudes above $2 * \sigma$ filtered out) and translation directions generated uniformly as points on the surface of a sphere. Rotations were generated in the same way, in the axis-angle representation, with the rotation amplitudes following the half-normal distribution ($\sigma = 0.3\pi$ rad, amplitudes above $2 * \sigma$ filtered out) and the rotation direction generated uniformly on the sphere. For each sample, the virtual C-arm camera was first rotated with the generated pose before deforming and rendering the CT volume (as described in Sec. 4.2.2). Examples of samples obtained with this process, as input to the network, are presented in Fig. 6.7.

The P020_small dataset was generated in the same way as the 'P020_p30' dataset, except that the standard deviations of the distributions were reduced to $\sigma = 50$ mm for the translation amplitudes, and $\sigma = 0.15\pi$ rad for the rotation amplitudes, with poses generated around the baseline pose only. Examples of samples obtained with this process, as input to the network, are presented in Fig. 6.8.

6.4.2 Results and discussion

We performed five experiments to test our method on this dataset.

In the first experiment, we trained the NoFix variant of our network on the P020_p30 dataset with a batch size of 12 rather than 7 and a learning rate of $5 \cdot 10^{-}3$, which are not the optimal parameters found above as this experiment was performed before the above experiments on the porcine dataset. We obtained a 3D target registration error of 6.6 mm $\pm$ 2.0 mm on the 80 test fluoroscopic images in the 30 deg range, from 5.08 mm $\pm$ 0 mm before registration. In 2D, we obtained a reprojection error of 9.9 mm $\pm$ 3.3 mm, from 7.4 mm $\pm$ 0.46 mm before registration.

Since the results were not satisfying, we tried to improve our method by leveraging more prior information about the content of the input image. To augment the network input, we used the lightweight variant of the MedSAM network developed by (Ma *et al.*, 2024), Ma *et al.* to automatically create 2D lung segmentation masks. MedSAM is a transformer neural network based on the SAM pretrained model (Kirillov *et al.*, 2023), fine-tuned on 1.5 million medical images distributed among ten different modalities. We further fine-tuned MedSAM on the P020_p30 dataset, which only contains 18 000 training samples, with as input a DRR image with a bounding box prompt covering the entire image, and, as a target, ground truth projected lung segmentations, using the loss combination proposed in the original work (Ma *et al.*, 2024). We then used the fine-tuned MedSAM model to create 2D lung segmentation masks for each sample in the P020_p30 dataset.

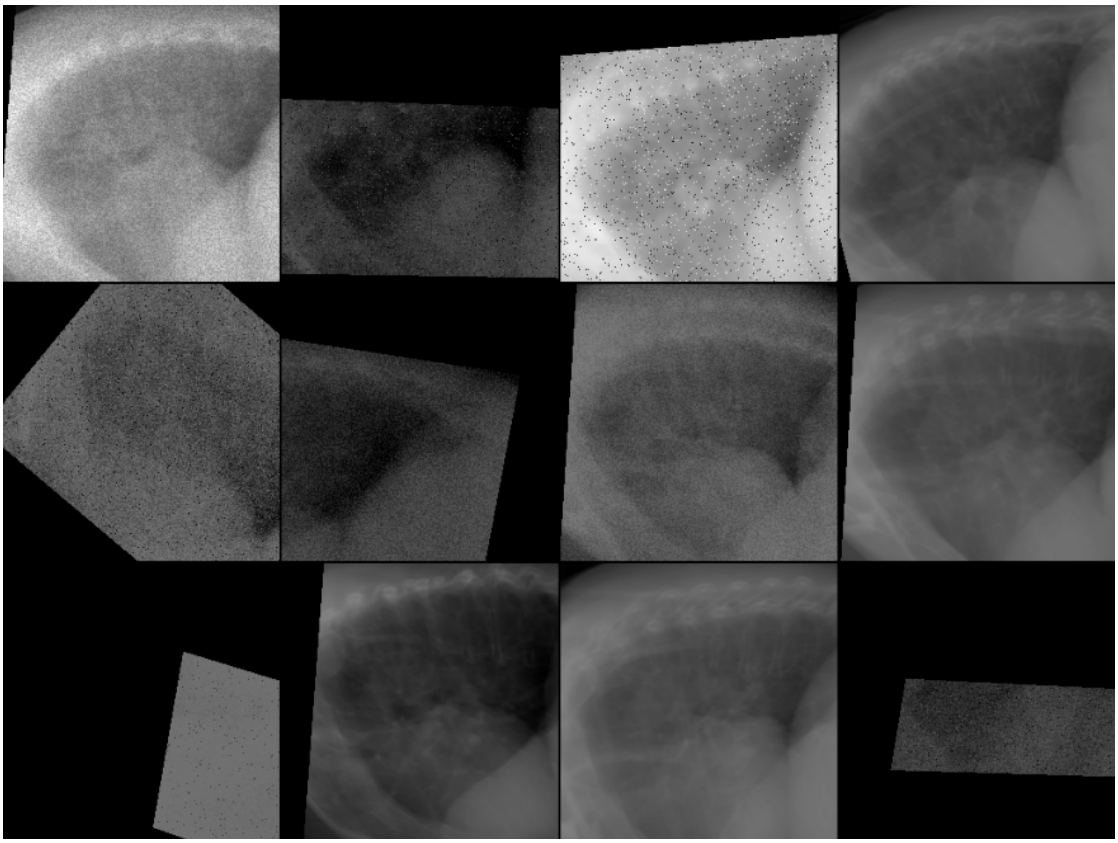


Figure 6.7: DRRs from the P020_p30 dataset. The DRRs are shown as input to the network after regions outside the field domain were masked and domain randomization was applied. Since regions outside the field domain show discontinuities in the displacement field, they must be masked when they appear in the projection, causing a loss of information in the image (see Sec. 4.2.5). Due to the large pose variations, some DRRs contain very little information (none in some cases) about the region of interest.

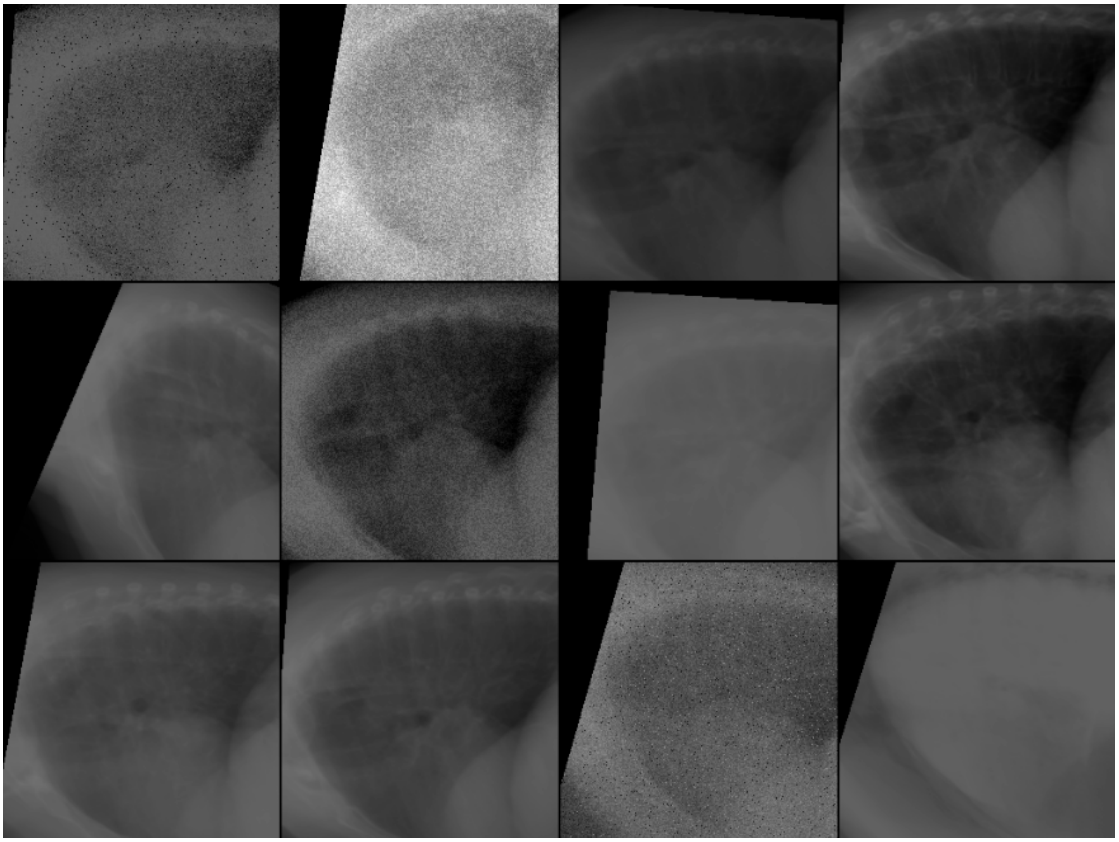


Figure 6.8: DRRs from the P020_small dataset, shown as input to the network. Thanks to the smaller pose variations, most image remain centered on the region interest (the right lung).

Thus, in our second experiment, we trained our network as in the first experiment, except for the network input which contained both a DRR and the lung segmentation mask predicted by MedSAM. After training, we obtained a 3D error of 7.4 mm ± 2.7 mm and a 2D error of 11.3 mm ± 4.6 mm.

In the third experiment, we trained the network once more with the lung segmentation mask as additional input, this time removing the masking operation after backprojection. Since the pose is changing in this dataset, as opposed to the porcine dataset, masking the features after backprojection may have a greater impact on the learning process, as it may remove too much information for the backpropagation process. In this experiment, we obtained improved results with a 3D error of 5.7 mm ± 2.4 mm and a 2D error of 7.4 mm ± 4.2 mm. The deformable registration performed by the network still does not improve on the prior registration, but this is the first experiment in which the network does not substantially degrade accuracy.

In the fourth experiment, we kept the same parameters as the previous experiment, except that we trained the network on the P020_small dataset instead, performing the same steps as before with MedSAM, including retraining, to generate input masks. In this experiment, we obtained a 3D error of 13.4 mm ± 2.7 mm and a 2D error of 21.9 mm ± 4.6 mm. Since the results were much worse than in previous experiments, we did not perform any further experiments on this dataset. It is possible that, in this dataset, the pose range did not cover the poses of test images, causing the network to fail.

Finally, in the last experiment, we used the optimal parameters found in Sec. 6.3.1 and did not use the segmentation mask in the input, as results were not clear on whether it improved performances or not. We obtained a 3D error of 10.1 mm ± 2.4 mm and a 2D error of 16.5 mm ± 4.0 mm. It is hard to interpret the results of this experiment, as the previously found optimal parameters do not seem to be effective here. It is also possible that the network architecture is not able to handle change of poses or that more training would be required for the network to learn this harder task.

The accuracy and parameters used in each experiment are summarized in table 6.1.

6.5 Conclusion

The experiments conducted on both porcine and clinical datasets provide valuable insights into the capabilities and limitations of the proposed registration method. In the porcine model experiments, the method achieved promising results with a median 2D error of 2.4 mm when using optimal parameters (10 training epochs and a batch size of 7). An implementation issue in the backprojection module was iden-

Experiment	Dataset	Input	Batch size	bp. mask	bp. fix	η	2D error (mm)	3D error (mm)
No reg.	NA	NA	NA	NA	NA	NA	7.4	5.1
Baseline	P020_p30	DRR	12	Yes	No	$5 \cdot 10^{-3}$	9.9	6.6
Input seg.	P020_p30	DRR + seg.	12	Yes	No	$5 \cdot 10^{-3}$	11.3	7.4
No bp mask	P020_p30	DRR + seg.	12	No	No	$5 \cdot 10^{-3}$	7.4	5.7
P020_small	P020_small	DRR + seg.	12	No	No	$5 \cdot 10^{-3}$	21.9	13.4
Optimal param.	P020_p30	DRR	7	No	Yes	$1 \cdot 10^{-4}$	16.5	10.1

Table 6.1: Accuracy obtained for each experiment on the clinical dataset. 'reg.' is short for registration, 'seg.' for segmentation, 'bp.' for backprojection, 'param.' for parameters and η refers to the learning rate.

tified and addressed, leading to improved performance in certain configurations (4 layers variant), but slightly decreased performances in others (10 layers variant), warranting further investigations. This experiment also highlighted the potential sensitivity of our method to unexpected appearance changes as evidenced by the consistent failure to register one marker positioned in a region showing changes in intestines content. Attempts to improve robustness through randomized intensity perturbations during training did not resolve this limitation, suggesting the alternate possibility of actual marker displacement rather than appearance-related issues.

The clinical experiments presented more significant challenges, particularly in handling the combined uncertainty on fluoroscopic image pose and CBCT to CT frame of reference change, rendering the experimental results hard to interpret. Using a pretrained 'foundation' model did not help to improve the results, despite its extensive training on real fluoroscopic image. Removing the 'masking' operation of features slightly helped, suggesting that the network might require more features to handle the combined rigid and deformable transformation between preoperative and intraoperative data. Our experiments on the P020_small dataset led to substantially worse performance, indicating the importance of maintaining sufficient pose variation during training. Finally, our last experiment using the optimal parameters found for the porcine dataset did not lead to improved performances, preventing us from drawing conclusions about the optimality of these parameters.

These findings highlight both the potential and current limitations of the method, particularly in bridging the gap between simulation and real clinical application. Future work should focus on creating a more controlled experimental testing dataset with available ground truth poses, for example through the use of a fluoroscopic registration phantom, as illustrated by Fig. 6.9. This next step will be particularly critical to help improve our method towards clinical use.

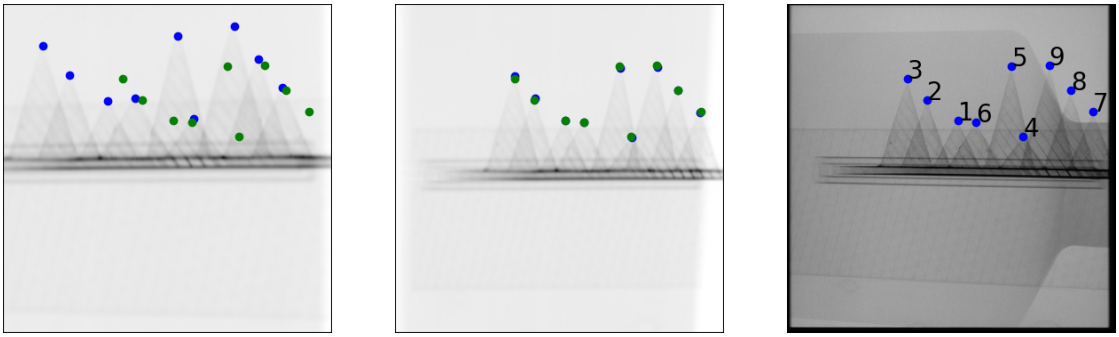


Figure 6.9: DRR of a 3D printed phantom at the initial camera pose (left), after pose optimization (middle) and corresponding fluoroscopic image (right).

Chapter 7

Conclusion & future directions

The increasing adoption of minimally invasive procedures has heightened the importance of intraoperative navigation solutions in clinical practice. While these procedures offer substantial patient benefits, they present unique challenges to clinicians, with significantly more complex visualization and manipulation during interventions.

Modern clinical workflows often integrate multiple imaging modalities, each serving distinct purposes throughout the intervention process. In the preoperative phase, three-dimensional modalities such as CT scans and MRI enable precise diagnosis and intervention planning, providing clinicians with detailed anatomical structure segmentation. During the intervention, real-time, 2D modalities are often preferred, due to their real-time capabilities.

Several approaches have emerged to harness the advantages of both 3D and 2D modalities, notably virtual bronchoscopy (section 1.2.2) and CBCT fluoroscopy (section 1.2.3). While these solutions have enhanced clinical practice, they retain certain limitations. Virtual bronchoscopy, though effective, remains confined to bronchoscopic procedures. CBCT fluoroscopy, while broadly applicable, exposes both patients and clinicians to higher radiation doses due to repeated scan acquisitions.

Conventional fluoroscopy provides real-time 2D anatomical visualization with relatively low radiation exposure. However, this modality exhibits poor contrast in certain organs, rendering some anatomical structures, like vessels, invisible. In current clinical practice, vessel visualization in fluoroscopic images is enabled by contrast agent injection. Yet, contrast agents are a suboptimal solution: they are nephrotoxic at high doses and provide only temporary visualization, as they dissipate with blood flow.

To address these limitations, we developed a versatile fluoroscopy-based interventional guidance technique. Our approach leverages preoperative CT data to enhance intraoperative fluoroscopic images through an innovative deep learning

framework. As detailed in section 4.6.2, we employ a patient-specific preoperative CT scan to train a deep neural network to recover deformations from fluoroscopic images. Our framework seamlessly integrates into existing clinical workflows, requiring minimal training time ($\approx$ one day) after preoperative CT scan acquisition, and no modifications to the interventional workflow.

Our methodology overcomes the limitations of existing 2D-3D deformable registration frameworks, which primarily target radiotherapy applications. Thanks to our domain-agnostic data generation framework (section 4.2.2), we are able to train a 2D-3D deformable registration deep neural network to recover arbitrary deformations. This capability renders our approach suitable to clinical contexts where intervention-related deformations occur, such as needle-based percutaneous procedures.

We validated our approach through comprehensive studies on both simulated (chapter 4) and real (chapter 6) fluoroscopic images. Preliminary experiments (section 4.3) demonstrated our domain-agnostic data generation's superiority over PCA-based methods, enabling our neural network to recover both arbitrary and breathing-induced deformations. A subsequent study (section 4.4) presents evidence of the clinical utility of our method through its ability to render vessels visible in fluoroscopic images, eliminating the need for contrast agent injection. We then improved, and more thoroughly evaluated our method in another study, presented in section 4.5, where we validated its ability to recover intervention-related deformations. The usefulness of our framework was further demonstrated through successful integration with an autonomous endovascular navigation system (section 4.6), significantly improving navigation success rates.

Chapter 5 explores the integration of biomechanical models to enhance the performance of our method. Such models, which have previously been used to model deformable organs in registration, employ physical parameters and partial shape information to predict complete organ configurations. Our initial investigation (section 5.2) examined the use of biomechanical models to generate physically accurate training deformations. Further experiments (section 5.3) explored biomechanical model-based regularization for ensuring physical plausibility. While initial attempts at developing a spatially differentiable network architecture (section 5.3.1) and implementing biomechanical model-based regularization (section 5.3.2) proved challenging, subsequent experiments with a simplified 2D registration problem demonstrated the potential of our biomechanical model-based regularizer in improving prediction realism and plausibility.

The final chapter 6 presents our method's evaluation on real fluoroscopic images. This semi-quantitative validation was difficult to set up, due to uncertainties on the ground truth 3D position of landmarks for the experiment on a porcine model, and on the pose of fluoroscopic images for the experiment on clinical data.

Nevertheless, the first experiment on the porcine model (section 6.3) showed that our method could successfully recover breathing motion for most landmarks, while the second experiment (6.4) on clinical data remained inconclusive.

To advance the clinical implementation of our method, a critical initial priority lies in conducting more comprehensive evaluation and enhancing the reliability of our network predictions on real fluoroscopic images. Our experimental findings in chapter 6 revealed that non-deformation related anatomical variations, such as changes in intestines content, may be responsible for inaccurate predictions. Despite our implementation of domain randomization techniques to address this challenge, our results remained inconclusive regarding their effectiveness.

A solution to this issue could be to refine our data generation process to simulate such anatomical changes. This could be achieved through automated organ segmentation followed by the generation of independent, organ-specific, deformations and intensity changes. Alternatively, we could enhance our data generation framework by leveraging large online databases of clinical CT scans to create a training dataset encompassing a broader spectrum of anatomical variations.

Modernizing our network architecture presents another avenue for improving prediction accuracy. Throughout this thesis work, we deliberately maintained our focus on convolutional networks, considering the rapid evolution of neural architectures and their increasing computational demands. However, recent developments in transformer architectures have demonstrated superior performance compared to convolutional neural networks when sufficient data and computational resources are available. The development of a large-scale 'foundation model' specifically designed for deformable 2D-3D registration could represent a significant advancement, requiring substantial initial investment, but potentially offering improved performance through subsequent fine-tuning for patient-specific and application-specific scenarios.

While our attempts to incorporate biomechanical model-based regularization for enhancing prediction realism and plausibility showed promise, our findings in chapter 5 indicate the need for further development in this direction. A promising approach would involve the integration of physical constraints directly within the network architecture, following the principles of 'physics-augmented neural networks' (PANNs) (Linden *et al.*, 2023). Such approach would also provide additional guarantees on the network's predictions, enhancing its reliability in a clinical setting.

Our experimental results have successfully demonstrated the capability of our approach to recover deformations from single fluoroscopic images in real-time, extending beyond the limitations of breathing motion-specific solutions. While multiple research directions remain to be explored for full clinical implementation, our current results provide strong evidence of our method's ability to effectively

recover deformations from clinical fluoroscopic images. These achievements establish a solid foundation for future developments while confirming the immediate practical utility of our approach.

Résumé

Introduction

Grâce au développement des techniques d'imagerie interventionnelle, les interventions mini-invasives, dans lesquelles l'opération est réalisée via des incisions de petite taille, se généralisent. De telles interventions sont désirables, car elles entraînent moins de complications et un temps d'hospitalisation réduit.

Dans une intervention guidée par imagerie, un scan 3D du patient est souvent acquis avant l'opération, à des fins de diagnostic et de planification. Pendant l'opération, des images sont acquises pour guider l'équipe clinique et leur permettre d'opérer de façon mini-invasive. Cependant, les modalités d'imagerie interventionnelles n'apportent qu'une information limitée, et ne permettent pas de voir en détail certaines structures anatomiques. À l'inverse, les modalités préopératoires comme le Computed Tomography (CT) scan permettent de localiser précisément, en 3D, la plupart des structures anatomiques.

Dans cette thèse, nous développons une solution pour améliorer les interventions guidées par fluoroscopie. Notre solution est basée sur le recalage déformable pour fusionner les informations préopératoire en 3D avec les images interventionnelles 2D. Grâce à cette fusion, les images interventionnelles peuvent être augmentées en temps réel avec des informations précises, mises à jour pour suivre les mouvements de l'anatomie. En effet, pendant l'intervention, l'anatomie est déformée par des mouvements anatomiques tels que la respiration, mais aussi par l'interaction des instruments chirurgicaux avec les tissus. Afin de corriger ces déformations et de permettre une fusion 2D-3D précise, une opération de recalage déformable est donc nécessaire.

Il existe des solutions, récemment développées et ayant fait l'objet d'études cliniques, pour augmenter l'information disponible dans les images interventionnelles. Cependant, celles-ci sont basées sur un recalage rigide ne prenant pas en compte les déformations, ou sur l'acquisition d'images 3D lors de l'intervention, qui ne permettent pas une visualisation en temps réel. Dans le cas particulier de la radiothérapie, des solutions de recalage déformable 2D-3D en temps réel ont été développées, mais elles ne prennent en compte que le mouvement respiratoire,

obtenu à partir de scans préopératoires 4D. Il n'existe donc pas de solution de recalage 2D-3D déformable pour les interventions guidées par fluoroscopie. Une telle solution éviterait en premier lieu l'acquisition de scans 3D interventionnels, réduisant l'exposition aux radiations et la durée de l'intervention. D'autre part, cette solution permettrait aussi de réduire ou supprimer l'injection d'agents de contraste pendant l'intervention, diminuant ainsi les risques de toxicité pour le patient tout en améliorant la visualisation de l'anatomie par l'équipe clinique.

Le premier chapitre de cette thèse introduit en détail le contexte clinique, présentant les avantages potentiels des interventions guidées par imagerie augmentées. Ce chapitre présente les solutions existantes pour augmenter les images interventionnelles, et plus particulièrement les solutions développées pour les interventions guidées par imagerie fluoroscopique.

Dans le second chapitre, le domaine scientifique du recalage d'image est présenté. En première partie de ce chapitre, il est fait état des différentes catégories de méthodes de recalage d'image et de leurs domaines d'application. La seconde partie présente brièvement le paradigme des méthodes de réseaux de neurones profonds appliqués au recalage d'image. Une courte présentation de l'historique des méthodes de réseaux de neurones profonds appliqués au recalage d'image 3D-3D, le cas le plus étudié dans la littérature, est proposée à la fin de ce chapitre.

Le troisième chapitre présente un état de l'art des méthodes de recalage déformable 2D-3D appliquées aux images fluoroscopiques, ainsi qu'un état de l'art des méthodes de recalage intégrant un modèle biomécanique.

Le quatrième chapitre présente le cœur de la méthode de recalage déformable développée durant cette thèse. Tout d'abord, le fonctionnement de la méthode et son insertion dans le processus interventionnel existant sont décrits en général. Notre méthode est basée sur l'utilisation du scan préopératoire pour la génération automatique de données d'entraînement. Ces données permettent d'entraîner de façon robuste, en un temps court, un réseau de neurone spécifiquement développé pour le recalage déformable 2D-3D. Le réseau de neurones entraîné est ensuite utilisé pour recaler, en temps réel, les données préopératoire sur les images interventionnelles. Le processus de génération de données d'entraînement, qui représente une contribution majeure de cette thèse, est tout d'abord décrit en détail. Ensuite, l'architecture du réseau de neurones pour le recalage déformable 2D-3D est présentée. La suite du chapitre présente les travaux effectués pour développer et valider notre approche. Parmi les travaux présentés dans ce chapitre, deux articles ont été publiés dans des conférences internationales et un article est en cours de relecture pour publication dans le journal Medical Image Analysis.

Le cinquième chapitre présente le travail effectué pour améliorer la méthode par l'utilisation d'un modèle d'organe biomécanique. L'objectif est ici de rendre le recalage plus réaliste en prenant en compte les lois de comportement biomé-

caniques des organes. Ce travail a donné lieu à la publication d'un article dans un 'Workshop' de la conférence internationale Computer Vision and Patter Recognition (CVPR).

Le sixième chapitre présente les expériences effectuées pour tester notre méthode sur des images fluoroscopiques expérimentales. De par les contraintes d'acquisition d'images, il est difficile d'évaluer quantitativement la précision du recalage déformable sur des images fluoroscopiques expérimentales. Les résultats présentés dans ce chapitre sont donc qualitatifs et semi-quantitatifs, mais permettent cependant d'évaluer l'efficacité de la méthode dans un contexte proche du contexte clinique.

Le dernier chapitre clôt ce travail de thèse et en résume les contributions. Un plan est également proposé, présentant les directions dans lesquels il serait intéressant de poursuivre le développement de la méthode pour s'approcher d'une utilisation clinique.

Le recalage d'images

Le recalage d'image est le processus qui établit une correspondance spatiale entre deux images. La transformation trouvée lors de ce processus permet de déformer l'image dite 'mobile' (I_M) pour la superposer à l'image dite 'fixe' (I_F). Cette technique est utile dans le domaine médical, où il est courant d'acquérir plusieurs images du même patient à différents points dans le temps. En effectuant le recalage entre ces images, il est par exemple possible d'étudier l'évolution d'une pathologie. La position du patient variant entre les sessions d'imagerie, un recalage rigide doit d'abord être effectué pour établir un alignement global entre les différentes acquisitions d'images. Cet alignement initial compense également les changements de référentiels entre les différents appareils d'imagerie, fournissant une base nécessaire pour les étapes de recalage ultérieures. En raison de processus physiologiques (respiration, mouvements cardiaques, ...), ou opératoires (action des cliniciens sur les organes), il est de plus nécessaire d'effectuer un recalage déformable pour aligner complètement les images.

Vue d'ensemble

Une première façon de catégoriser les méthodes de recalage est de les séparer en deux catégories : les méthodes de recalage unimodales et les méthodes de recalage multimodales. Dans le cas du recalage unimodal, la modalité d'imagerie est la même pour toutes les images, comme par exemple pour les méthodes développées pour recaler des scans CT. À l'inverse, dans le cas du recalage multimodal, les images sont acquises dans des modalités différentes. Notre méthode s'inscrit dans

cette catégorie en recalant un scan CT avec une image fluoroscopique, un problème plus difficile à résoudre que le recalage entre deux images de même modalité et de même dimensionalités, mais combinant les avantages des deux modalités pour visualiser sur les images fluoroscopiques les informations issues du CT.

Il est également possible de distinguer les méthodes de recalage en fonction de la nature de la transformation utilisée, qui peut être rigide ou déformable. Bien que recherchant une transformation plus simple (affine), le recalage rigide est une première étape indispensable dans le processus de recalage. Le recalage déformable, quant à lui, prend en compte les déformations de l'anatomie, à travers l'utilisation de transformation non-linéaires. Ces transformations peuvent être définies à l'aide de paramètres ou bien par un champ de déplacement dense.

Une catégorie particulière de méthodes paramétriques utilise des modèles biomécaniques pour guider le recalage, par la prise en compte des paramètres physiques de l'organe. Ces modèles contraignent le recalage à respecter les lois de la mécanique, permettant ainsi d'extrapoler de façon réaliste la déformation dans les parties non observées de l'organe. Le modèle biomécanique d'élasticité linéaire est simple mais inexact pour les grandes déformations, motivant l'utilisation de modèles hyperélastiques, plus complexes mais plus précis. Les modèles biomécaniques, décrits par des équations différentielles, nécessitent l'utilisation de méthodes numériques telles que la méthode des éléments finis (MEF) ou la méthode des différences finies (MDF). Ces méthodes, bien que précises, sont coûteuses en temps de calcul, nécessitant l'utilisation d'approximations pour les applications temps réel. Malgré ces contraintes, les méthodes biomécaniques ont été utilisées avec succès pour le recalage d'images médicales.

Les méthodes de recalage se distinguent également entre approches basées sur l'intensité et approches basées sur les caractéristiques. Les premières utilisent les intensités des pixels dans les images, tandis que les secondes utilisent des caractéristiques extraites des images, telles que des maillages d'organes. Ces dernières, bien que plus robustes aux différences entre les images, nécessitent une étape non-triviale de pré-traitement pour extraire les caractéristiques.

Un point commun entre la plupart des méthodes de recalage est l'utilisation d'une fonction de coût L minimisée par un algorithme d'optimisation pour trouver la transformation de recalage T. Traditionnellement, l'algorithme d'optimisation est utilisé pour effectuer le recalage entre deux images, nécessitant dans certains cas plusieurs heures de calcul. Pour pallier cette limitation, les méthodes d'apprentissage cherchent à minimiser la fonction de coût sur une base de données d'images, calculant à travers ce processus les paramètres θ d'une fonction f_θ qui associe T à (I_M, I_F). Ainsi, une fois le processus d'apprentissage terminé, f_θ permet de calculer rapidement la transformation de recalage entre deux images.

Les méthodes basées sur Modèle de Déformation Statistique (MDS) constituent

un exemple de méthodes d'apprentissage largement utilisées pour le recalage. Le fonctionnement général des méthodes basées sur un MDS est le suivant :

1. À partir d'un ensemble de N images, on définit une image fixe de référence associée à un ensemble d'images à recaler.

2. Après avoir calculé la transformation de recalage entre l'image de référence et chaque image de l'ensemble, on obtient un ensemble de transformations T_i pour $i \in \{1, \ldots, N\}$.

3. Ensuite, l'Analyse en Composantes Principales (ACP) est souvent utilisée pour représenter T_i avec k composantes principales.

4. En pratique, k est petit, avec par exemple $k = 3$ pour un mouvement respiratoire.

5. Pour recaler une nouvelle image avec l'image de référence, il suffit d'optimiser k paramètres pour obtenir la transformation de recalage, fonction des k composantes principales.

La limitation des méthodes basées sur un MDS est qu'elles ne permettent pas de retrouver une transformation qui ne soit pas une combinaison linéaire des transformations de l'ensemble d'apprentissage. Cette contrainte restreint très fortement les transformations possibles, ce qui est utile pour retrouver des déformations connues a priori (respiration, mouvement cardiaque), mais ne permet pas de retrouver des transformations arbitraires (déformations causées par l'action d'un instrument chirurgical).

L'apprentissage profond pour le recalage d'images

Pour surmonter les limitations des méthodes de recalage traditionnelles, les méthodes basées sur l'apprentissage profond ont récemment été utilisées de façon massive. L'apprentissage profond est une approche d'apprentissage automatique, dans laquelle les paramètres d'un réseau de neurones artificiel sont optimisés sur une base de données. Un réseau de neurones artificiel est une composition de fonctions paramétriques et non linéaires. Il est généralement composé d'au moins trois couches, chacune formée par la composition d'une fonction paramétrique et d'une fonction d'activation non linéaire (et éventuellement de fonctions supplémentaires). Ces réseaux, qui permettent en théorie d'approximer n'importe quelle fonction, ont permis de résoudre de nombreux problèmes complexes avec une efficacité supérieure aux méthodes traditionnelles. Ils sont généralement entraînés à l'aide de la différentiation automatique, qui permet l'utilisation de l'algorithme de propagation du gradient pour trouver les paramètres θ du réseau minimisant

L. Cette méthode, basée sur le calcul des dérivées, est étonnamment efficace lorsqu'une grande quantité de données est disponible, permettant l'utilisation de réseaux pouvant avoir des milliards de paramètres, disposant d'une capacité de représentation jusqu'ici inégalée. Cette capacité de représentation est particulièrement utile pour le recalage d'images, où la transformation f_θ associant T à (I_M, I_F) est potentiellement de très grande complexité.

Les premières méthodes d'apprentissage profond pour le recalage d'images ont été proposées dans les années 2010. La plupart de ces méthodes ont été développées pour le recalage unimodal d'images médicales 3D, n'apportant pas nécessairement des performances supérieures, mais accélérant le processus de recalage jusqu'à 20 000 fois par rapport aux méthodes itératives traditionnelles, pour une précision équivalente. Cette caractéristiques les rend particulièrement intéressantes pour notre application, le recalage déformable 2D-3D en temps réel.

État de l'art du recalage déformable 2D-3D

Les modalités d'imagerie 2D sont utilisées dans les interventions en raison de leurs capacité à fournir une visualisation en temps réel de l'anatomie. La fluoroscopie est une modalité 2D qui présente l'avantage d'offrir un large champ de vision et de montrer l'anatomie interne complète du patient. Elle est utilisée notamment pour suivre les cathéters lors des procédures endovasculaires, visualiser les aiguilles dans les procédures percutanées et guider les opérations orthopédiques.

Une limitation majeure des images fluoroscopiques est le manque de contraste entre les tissus de densité similaire, ce qui rend difficile la distinction des structures anatomiques. Pour pallier cette limitation, des agents de contraste peuvent être injectés ou des marqueurs radio-opaques implantés, mais, comme évoqué précédemment, ces solutions ne sont pas idéales.

Pour apporter une visualisation de structures anatomiques en continu sans risques pour le patient, les méthodes de recalage déformable 2D-3D superposent des informations issues d'une image 3D préopératoire à une image 2D peropératoire en temps réel. Nous nous plaçons dans le contexte clinique où l'image préopératoire est un scan CT et l'image peropératoire est une image fluoroscopique.

Il convient tout d'abord de noter que le recalage déformable n'est pas la seule façon d'améliorer la visualisation de structures anatomiques sur des images fluoroscopiques. En effet, des méthodes de localisation sans marqueurs existent également pour le traitement de tumeurs par radiothérapie (Shieh *et al.*, 2017; Hirai *et al.*, 2019; W. Zhao *et al.*, 2019; Zhang, X. Huang, Wang, *et al.*, 2020; Y. Yan *et al.*, 2024). Cependant, ces méthodes ne permettent pas d'incorporer d'informations préopératoires et sont donc moins versatiles que les méthodes de recalage.

Une autre catégorie de méthodes de recalage 2D-3D concerne l'estimation de

pose, ou recalage rigide, adaptée à la visualisation de structures rigides, telles que les structures osseuses (D. C. Liu *et al.*, 1989; Berger *et al.*, 2016; Hansen *et al.*, 2003; Powell *et al.*, 2009; Wunsch *et al.*, 1996; Benameur *et al.*, 2003; Gall *et al.*, 1993; T. S. Tang *et al.*, 2000; Gouveia *et al.*, 2012; D.-X. Huang *et al.*, 2024; Jaganathan *et al.*, 2023; Gao, Killeen, *et al.*, 2023; Gao, Feng, *et al.*, 2023; Gopalakrishnan, Dey, *et al.*, 2024; M. Chen, Z. Zhang, Gu, Ge, *et al.*, 2024; M. Chen, Z. Zhang, Gu, and Kong, 2024; B. C. Lee *et al.*, 2022). Ces méthodes sont également nécessaires en tant qu'étape préliminaire au recalage déformable, et montrent une précision souvent inférieure au millimètre, mais échouent parfois à converger vers un recalage correct.

La majorité des méthodes de recalage déformable 2D-3D dans la littérature se concentrent sur la compensation des mouvements respiratoires ou cardiaques (C.-R. Chou, Frederick, *et al.*, 2013; C.-R. Chou and Pizer, 2013; C.-R. Chou, Frederick, *et al.*, 2013; M. D. Foote *et al.*, 2019; Nakao *et al.*, 2022; Shao, Jing Wang, *et al.*, 2022; Wijesinghe, 2024). Ces méthodes utilisent pour la plupart un MDS issu d'un CT 4D préopératoire ou d'une base de données de CT 4D. Cette limitation à l'estimation de mouvement périodiques à partir de données 4D rend ces méthodes peu adaptées au recalage déformable 2D-3D dans le cadre de déformations liées à l'action des cliniciens sur les organes.

Afin d'améliorer la précision de notre méthode, nous nous intéressons également aux approches basées sur des modèles biomécaniques. Ces approches utilisent un modèle élastique (Broit, 1981) ou hyperélastique (Rabbitt *et al.*, 1995; Pennec *et al.*, 2005; Yanovsky *et al.*, 2008; P. Alvarez, Rouzé, *et al.*, 2021; Lesage *et al.*, 2020) pour modéliser la déformation de l'organe. Parmi ces méthodes, certaines ont été développées pour le recalage d'images CBCT interventionnelles du poumon (P. Alvarez, Rouzé, *et al.*, 2021; Lesage *et al.*, 2020), qui subit de grandes déformations dues au phénomène de pneumothorax lors de l'incision du thorax.

Recalage déformable 2D-3D agnostique au domaine

Afin de développer une méthode de recalage déformable 2D-3D adaptée aux interventions guidées par fluoroscopie, notre méthode s'écarte des méthodes existantes, qui utilisent un MDS pour la génération des données d'apprentissage. S'inspirant des méthodes de reconstruction de volume CT à partir d'images fluoroscopiques (Shen *et al.*, 2019; Yikun Zhang *et al.*, 2021; J. Guo *et al.*, 2024), notre approche se base sur les résultats de Shen *et al.* (Shen *et al.*, 2019), qui ont montré qu'il était possible de reconstruire un volume 3D à partir d'une seule image fluoroscopique, avec une précision limitée.

Le réseau de neurones utilisé dans la méthode de Shen *et al.* est un réseau de neurones convolutif (RNC) utilisant l'architecture ResNet (K. He *et al.*, 2016).

Dans cette architecture, des connections directes sont ajoutées entre les couches du réseau afin de faciliter la propagation des gradients et donc l'apprentissage. Pour prédire un volume 3D à partir d'une image fluoroscopique 2D, l'image fluoroscopique en entrée est d'abord transformée par l'encodeur du réseau, composée de 11 couches de convolution, en un tenseur de 'feature maps', une représentation abstraite de l'image de faible résolution spatiale, riche en informations. Ces 'feature maps' sont ensuite transformées de 2D en 3D par un module de transformation qui ne contient pas de paramètres mais réordonne l'ensemble de 'feature maps' 2D pour former un ensemble réduit de 'feature maps' 3D. Finalement, le décodeur du réseau, composé de 11 couches de convolution, transforme ces 'feature maps' 3D en un volume 3D de haute résolution spatiale.

Cette architecture présente l'avantage de la simplicité en transformant directement une image fluoroscopique 2D en un volume CT 3D grâce au module de transformation 2D-3D. L'architecture que nous utilisons est basée sur l'architecture proposée par Shen *et al.*, avec quelques différences. Comme l'objectif de notre approche est de superposer les informations issues du volume préopératoire avec l'image fluoroscopique plutôt que de reconstruire un volume 3D, nous avons modifié la dernière couche du décodeur pour produire un champ de déplacement 3D au lieu d'un volume 3D, cette approche ayant déjà été employée par des méthodes de recalage 3D (De Vos *et al.*, 2017; Shan *et al.*, 2017; Miao *et al.*, 2018; Balakrishnan *et al.*, 2019; J. Chen *et al.*, 2022; Y. Zhu *et al.*, 2022).

Etant donné le manque de base de données réelles pour le recalage déformable 2D-3D, contenant des paires d'images fluoroscopiques et de volumes CT, nous avons développé un processus de génération de données synthétiques. A partir d'un scan CT préopératoire auquel nous appliquons des déformations aléatoires, nous obtenons une base de données d'images fluoroscopiques synthétiques contenant des déformations. Cette approche permet d'entraîner le réseau de neurones à retrouver une déformation du CT à partir d'une image fluoroscopique.

La section suivante présente notre méthode de recalage déformable 2D-3D, évaluée à travers plusieurs études. Notre principale contribution est notre approche de génération de données agnostique au domaine, qui permet de s'affranchir des connaissances préalables sur les mouvements pendant les interventions guidées par fluoroscopie. Nous avons également amélioré l'architecture proposée par Shen *et al.* en utilisant un module de rétroprojection pour transformer les 'feature maps' 2D en 'feature maps' 3D en prenant en compte les informations de pose. Nous avons aussi introduit une nouvelle fonction de coût pour superviser le réseau dans l'espace projectif, prenant en compte la perte d'informations induite par la projection. Les études présentées démontrent la supériorité de cette approche par rapport aux méthodes basées sur un MDS pour la prédiction de déformations non périodiques, et son potentiel pour réduire l'utilisation d'agents de contraste lors

des interventions percutanées.

Méthode développée

Notre méthode s'intègre dans la pratique clinique existante tel qu'illustré par la Fig. 7.1.

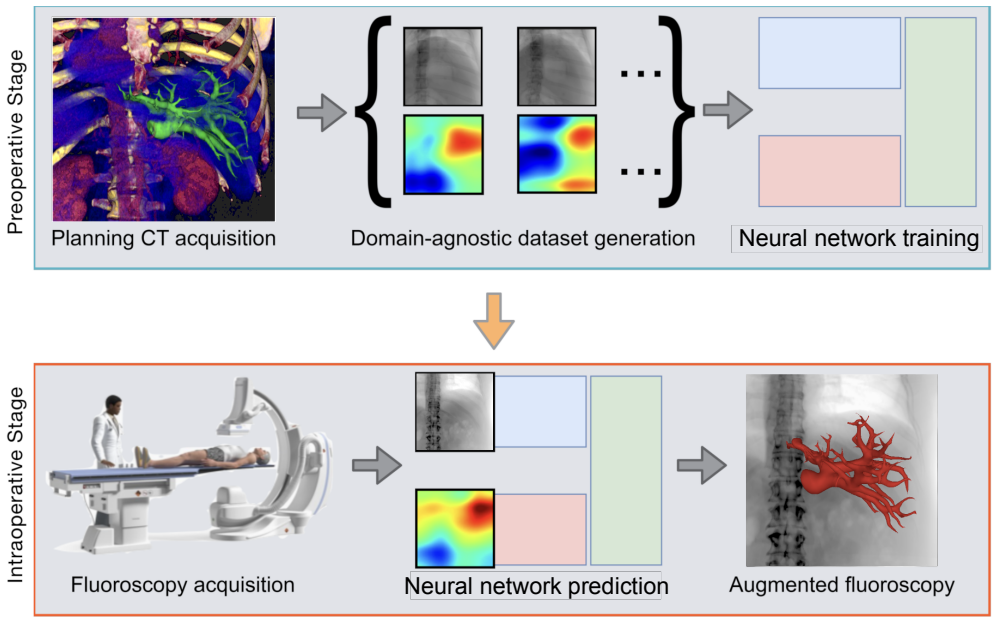


Figure 7.1: Présentation de notre approche. La première étape est la planification de l'intervention à partir d'un scan CT préopératoire, où les structures d'intérêt sont segmentées et la pose du C-arm est déterminée (en haut à gauche). La seconde étape est l'entraînement du réseau de neurones à prédire une déformation du CT 3D à partir d'une image fluoroscopique synthétique (en haut, au milieu et à droite). Ici, les déformations sont représentées schématiquement en 2D (en réalité, les déformations sont des champs vectoriels 3D), la couleur indiquant l'amplitude du déplacement. Ensuite, lors de l'intervention, le C-arm est positionné et une image fluoroscopique est acquise (en bas à gauche). Enfin, le réseau de neurones est utilisé pour prédire la déformation du CT 3D à partir de l'image fluoroscopique (en bas, au milieu), permettant d'obtenir une image fluoroscopique augmentée (en bas, à droite).

Le processus de génération de données est basée sur l'utilisation du scan CT préopératoire, éventuellement injecté pour segmenter les vaisseaux. La pose du C-arm par rapport au CT préopératoire est également supposée connue, comme dans d'autres approches de la littérature. En pratique, cette hypothèse n'est pas vérifiée,

mais il est possible de combiner notre approche avec une méthode de recalage rigide 2D-3D pour obtenir la pose intraopératoire, et d'autre part, d'introduire des variations de pose dans le processus de génération de données pour se passer de cette hypothèse. Néanmoins, comme il s'agit ici de démontrer la faisabilité du recalage déformable 2D-3D pour les interventions guidées par fluoroscopie, nous conservons dans un premier temps cette hypothèse.

Tout d'abord, le scan CT préopératoire injecté est traité pour enlever les informations de contraste des vaisseaux en remplaçant les voxels autours des vaisseaux par des voxels d'intensité moyenne dans l'organe ('inpainting'). Ensuite, une base de données d'entraînement est générée en appliquant des déformations au CT préopératoire, puis en générant des images fluoroscopiques synthétiques à partir des CT déformés. Ces données sont utilisées pour entraîner le réseau à prédire la déformation du CT à partir de l'image fluoroscopique synthétique, dans le but d'augmenter, lors de l'intervention, les images fluoroscopiques avec les informations issues du CT préopératoire.

Pour générer des déformations agnostiques au domaine, nous cherchons à générer aléatoirement des déformations lisses et inversibles, deux propriétés vérifiées pour les déformations réelles. Le 'Large Deformation Diffeomorphic Metric Mapping (LDDMM)' (Trouve *et al.*, 2005) est une méthode développée pour produire de telles déformations dans le cadre du recalage d'images. Dans le LDDMM, la déformation $\boldsymbol{\phi}$ qui recale une image I vers une image I' est obtenue en intégrant un champ de vitesse $\mathbf{V}(t, \mathbf{x})$ gouverné par un ensemble d'équations différentielles. Il a été démontré (Durrleman *et al.*, 2014) qu'il était possible d'exprimer $\mathbf{V}(t, \mathbf{x})$ par l'équation suivante :

$$\mathbf{V}(t, \mathbf{x}) = \sum_{k=1}^{N_{cp}} \boldsymbol{\alpha_k}(\boldsymbol{t}) \cdot \mathbf{K_k}(\mathbf{x}, \mathbf{y_k}(\mathbf{t})) \tag{7.1}$$

où $\mathbf{K_k}(t)$ sont des éléments d'un espace de Hilbert à noyau reproduisant. Il est donc possible d'utiliser des noyaux gaussiens pour représenter $\mathbf{K_k}(t)$, situés aux N_{cp} points de contrôle $\mathbf{y_k} \in \mathbb{R}^3$ et pondérés par les coefficients $\boldsymbol{\alpha_k}(t) \in \mathbb{R}^3$. Le champ de déplacement est alors obtenu en calculant $\boldsymbol{\varphi}(\mathbf{x}) = \int_0^1 \mathbf{V}(t, \mathbf{x})dt$.

Plutôt que d'optimiser les paramètres $\boldsymbol{\alpha_k}$ et $\mathbf{y_k}$ pour le recalage, nous générons des valeurs aléatoires des points de contrôle $\mathbf{y_k}$, matrices de covariance $\boldsymbol{\sigma_k} \in \mathbb{R}^{3\times3}$ et poids $\boldsymbol{\alpha_k}$ des noyaux gaussiens. Les points $\mathbf{y_k}$ sont d'abord générés aléatoirement puis filtrés pour ne garder que les points conservant une distance minimale, afin d'éviter d'obtenir des noyaux trop proches les uns des autres, ce qui pourrait entraîner des variations trop rapides de $\boldsymbol{\varphi}(\mathbf{x})$. Les coefficients $\boldsymbol{\alpha_k}$ sont générés à partir d'une distribution uniforme sphérique en 3D puis multipliés par un facteur commun entre -1 et 1 pour garantir que des champs de déplacement faibles partout seront générés. Finalement, les paramètres $\boldsymbol{\sigma_k}$ sont générés de façon uniforme avec

des valeurs entre 15% and 30% de la taille du volume. Bien que cette formulation ne garantisse pas que $\boldsymbol{\phi}$ soit difféomorphique, on vérifie en pratique que la valeur du jacobien spatial J est positive pour garantir le difféomorphisme.

Pour générer des images fluoroscopiques synthétiques, nous utilisons l'algorithme DeepDRR (Unberath *et al.*, 2018), qui modélise le C-arm comme une caméra utilisant des rayons X pour former des images. A partir d'un scan CT $I(\mathbf{x})$, l'image fluoroscopique synthétique p est obtenue par l'équation suivante :

$$p(\mathbf{u}) \approx \int I(\mathbf{x}) \mathbf{dl}_{\mathbf{u}} \tag{7.2}$$

avec $\mathbf{l}_{\mathbf{u}}(\mathbf{x}) = \mathbf{P} \cdot \mathbf{x}$ le rayon qui connecte le point $\mathbf{u} \in \mathbb{R}^2$ sur le plan de détection à la source de rayons X, P la matrice de projection et $I(\mathbf{x})$ l'intensité du scan CT à la position $\mathbf{x} \in \mathbb{R}^3$. Cette équation montre que $p(\mathbf{u})$ est invariant aux transformations de $I(\mathbf{x})$ qui préservent la valeur de l'intégrale. Cela signifie qu'une déformation déplaçant des voxels le long d'un rayon de projection ne modifie pas la valeur de l'intégrale, rendant une telle déformation inobservable dans l'image.

L'architecture de notre réseau, comme évoqué plus haut, est dérivée de (Shen *et al.*, 2019), et est représentée par la figure 7.2. Après le calcul des 'feature maps' par les 10 couches de convolution, celles-ci sont transformées en 3D par le module de transformation 2D-3D. En prenant en compte la pose P de l'image fluoroscopique par rapport au scan CT, ce module permet de faire correspondre le champ de déplacement à la position $\mathbf{x}$ dans le référentiel du CT aux informations extraites du pixel $\mathbf{u}$ à la position $\mathbf{u} = \mathbf{Px}$ dans l'image fluoroscopique. Le module de transformation 2D-3D ne nécessite pas de paramètres à apprendre et utilise une grille de points en 3D $\mathbf{G}$ pour interpoler la valeur des feature maps aux positions 2D correspondantes $\mathbf{G_u} = \mathbf{PG}$. Dans ce module, les 'feature maps' 2D sont considérées comme une succession de plans image à différentes profondeurs le long des lignes de projection. Après rétroprojection en 3D, on obtient un volume 3D formé par interpolation de feature maps 2D dans le cône de projection, avec des valeurs nulles en dehors. Pour obtenir des 'feature maps' 3D à partir de ce volume, la dimension de profondeur est divisée en deux pour obtenir plusieurs volumes, formant les feature maps. Finalement, ces 'feature maps' sont traitées par le décodeur pour obtenir la prédiction du champ de déplacement.

Pour entraîner le réseau, nous utilisons une combinaison $\mathcal{L}$ de fonctions de coût, en prenant soin de masquer les voxels non visibles dans l'image fluoroscopique ou en dehors de l'anatomie.

$$\mathcal{L} = \mathcal{L}_{\boldsymbol{\varphi}^{2D}}(\boldsymbol{\varphi}_{\boldsymbol{i}}^{2D}, \hat{\boldsymbol{\varphi}}_{\boldsymbol{i}}{}^{2D}) + \lambda \mathcal{L}_{s^{2D}}(\mathcal{P}_{2D}(I_s \circ \boldsymbol{\phi}_{\boldsymbol{i}}), \mathcal{P}_{2D}(I_s \circ \hat{\boldsymbol{\phi}}_{\boldsymbol{i}})) \tag{7.3}$$

Le premier terme de cette combinaison, $\mathcal{L}_{\boldsymbol{\varphi}^{2D}}$, pénalise l'erreur sur le déplacement dans l'espace 2D plutôt que dans l'espace 3D. En effet, tous les déplacements ne

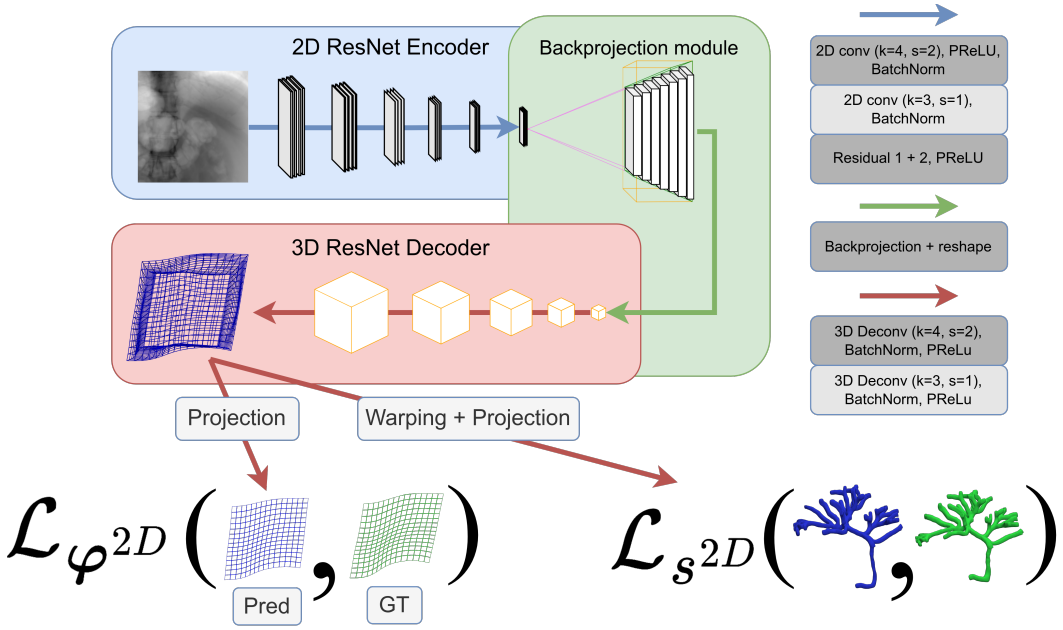


Figure 7.2: L'encodeur du réseau convertit l'image en un ensemble de 'feature maps' de basse résolution spatiales. Les 'features maps' 2D sont rétroprojetées en 3D à l'aide de la matrice de projection dans le module de rétroprojection. Le décodeur du réseau convertit les 'feature maps' 3D en un champ de déplacement. Le champ de déplacement prédit est ensuite utilisé pour calculer la fonction de coût sur les points de grille projetés dans l'espace 2D et sur la segmentation des vaisseaux projetés dans l'espace 2D.

sont pas observables dans l'image fluoroscopique, et ne peuvent donc pas être prédit sans informations supplémentaires. De plus, cette fonction de coût est naturellement adaptée à notre objectif, la visualisation en 2D d'informations anatomiques 3D, dont la projection doit être exacte. Le deuxième terme, $\mathcal{L}_{s^{2D}}$, permet de minimiser l'erreur en 2D au niveau des structures anatomiques d'intérêt, en utilisant le coefficient de Dice entre les segmentations déformées prédites et de référence.

Pour améliorer la robustesse du réseau aux changements d'apparence entre les images fluoroscopiques synthétiques d'entraînement et les images fluoroscopiques réelles, nous utilisons une méthode d'augmentation de données décrite dans la section 'post-processing' de (Grimm *et al.*, 2021). Cette méthode consiste à ajouter différentes sortes de bruit de façon aléatoire aux images d'entraînement, rendant ainsi le réseau plus robuste aux changements d'apparence.

Le processus d'entraînement nécessite de plus diverses transformations des données. Lors de l'initialisation, nous chargeons les paramètres de projection DRR, la géométrie du volume CT, la segmentation utilisée pour $\mathcal{L}_{s^{2D}}$. La transforma-

tion appliquée aux images en entrée est également définie, ainsi que les variables dépendantes de la pose, utilisées à chaque itération. Cette transformation inclut l'application d'un masque dépendant de la pose, servant à exclure les régions de l'image montrant les limites du volume CT. Ce masque en 2D permet de calculer le masque en 3D, évoqué ci-dessus, qui définit les voxels visibles dans l'image fluoroscopique.

Ensuite, au début de chaque itération, la transformation définie plus haut est appliquée aux images en entrée et le masque 3D correspondant à l'intérieur de l'anatomie est mis à jour avec le champ de déplacement de référence.

Enfin, après la prédiction, nous appliquons le masque 3D aux champs de déplacement, déformons puis effectuons le rendu des segmentations en 2D avec une technique de lancer de rayons adaptée de DiffDRR (Gopalakrishnan and Golland, 2022). Finalement, les champs de déplacement sont convertis en millimètres et le masque utilisé pour l'image d'entrée est appliqué aux segmentations projetées s^{2D} avant le calcul de $\mathcal{L}$.

Afin d'évaluer la précision de notre méthode, nous avons réalisé plusieurs études à partir de scans CT de patients ou de modèles porcins. Dans ces expériences, le réseau est d'abord entraîné en générant des images fluoroscopiques synthétiques à partir d'un volume CT préopératoire selon le processus décrit précédemment. Puis, la précision du réseau est mesurée sur une ou plusieurs images fluoroscopiques synthétiques générées à partir d'un volume CT différent du même patient ou modèle porcin, pour lequel la position des structures anatomiques est connue.

Étude préliminaire

Dans une étude préliminaire, nous avons comparé notre méthode de génération de données avec une méthode de génération de données basée sur un MDS. Le réseau utilisé dans cette étude n'intègre pas encore le module de rétroprojection proposé, et utilise une fonction de coût sur le déplacement 3D, masquant la composante du champ de déplacement parallèle à l'axe de la caméra. D'autres différences concernent également le processus de génération et d'augmentation de données, ainsi que les résolutions spatiales utilisées et le processus d'entraînement.

Cette étude utilise un volume CT 4D issu de la base de données présentée dans (Hugo *et al.*, 2017) et montrant un mouvement respiratoire à travers 10 volumes 3D. Le premier volume est utilisé pour générer les données d'entraînement agnostiques au domaine, tandis que les 9 autres sont utilisés pour évaluer la précision du réseau. Le MDS, d'autre part, est généré à partir des 9 volumes CT en les recalant au premier volume CT avec l'algorithme de recalage 3D SyN (B. B. Avants *et al.*, 2008) puis en extrayant les trois composantes principales du mouvement par ACP. En variant la valeur des coefficients de ces composantes, des déformations proches du mouvement respiratoire sont générées pour entraîner le

réseau. Comme dans d'autres études utilisant un MDS, comme par exemple (M. D. Foote *et al.*, 2019) réalisée sur la même base de données, cette approche présente un biais puisqu'elle teste le réseau sur le même mouvement respiratoire que celui utilisé pour générer les données d'entraînement.

Malgré ce biais défavorable à notre approche agnostique au domaine, qui n'utilise pas de MDS, nous obtenons des résultats comparables avec ceux de (M. D. Foote *et al.*, 2019) en testant sur des images fluoroscopiques issues du CT 4D (mouvement respiratoire). Nous obtenons également une erreur maximale améliorée de $2,22$ mm ($9,55$ mm dans (M. D. Foote *et al.*, 2019)).

Cette étude préliminaire nous a permis de valider notre approche de génération de données en principe et de montrer qu'il n'était pas nécessaire d'utiliser un MDS, et donc un CT 4D, pour entraîner un réseau pour le recalage déformable 2D-3D.

Visualisation des vaisseaux sans contraste

Dans une seconde étude, présentée à la conférence internationale 'Hamlyn Symposium on Medical Robotics 2023' (HSMR), nous avons évalué la capacité de notre méthode à remplacer l'injection d'agents de contraste pour la visualisation des vaisseaux du foie dans les images fluoroscopiques :

> Francois Lecomte *et al.* (2023). "Enhancing fluoroscopy-guided interventions: a neural network to predict vessel deformation without contrast agents". In: *The Hamlyn Symposium on Medical Robotics*. The Hamlyn Centre, Imperial College London London, UK, pp. 75–76

Les données d'entraînement sont générées de la même manière que dans l'étude précédente, à partir d'un volume CT préopératoire d'un patient de l'hôpital Paul Brousse à Paris, dans lequel les veines hépatiques ont été segmentées. L'architecture du réseau présente des différences par rapport à l'étude précédente, mais la fonction de coût est la même. Comme un CT 4D n'est pas disponible dans ces données cliniques, nous avons créé un mouvement respiratoire synthétique pour générer les données de test. Nous avons mesuré la précision du réseau sur un nuage de points issu de la segmentation des veines hépatiques dans l'espace projectif 2D. L'amplitude moyenne du mouvement respiratoire en 2D est de $7,7 \pm 3,9$ mm, tandis que la précision moyenne du réseau est de $2,7 \pm 1,9$ mm, validant la capacité de notre méthode à retrouver la position des veines hépatiques en 2D avec une précision < 3 mm lors du mouvement respiratoire. Une vidéo montrant les résultats de cette étude est disponible à cette adresse : `https://mimesis.inria.fr/project/augmented-fluoroscopy/`. La figure 7.3 montre la visualisation obtenue avec notre méthode lors du mouvement respiratoire.

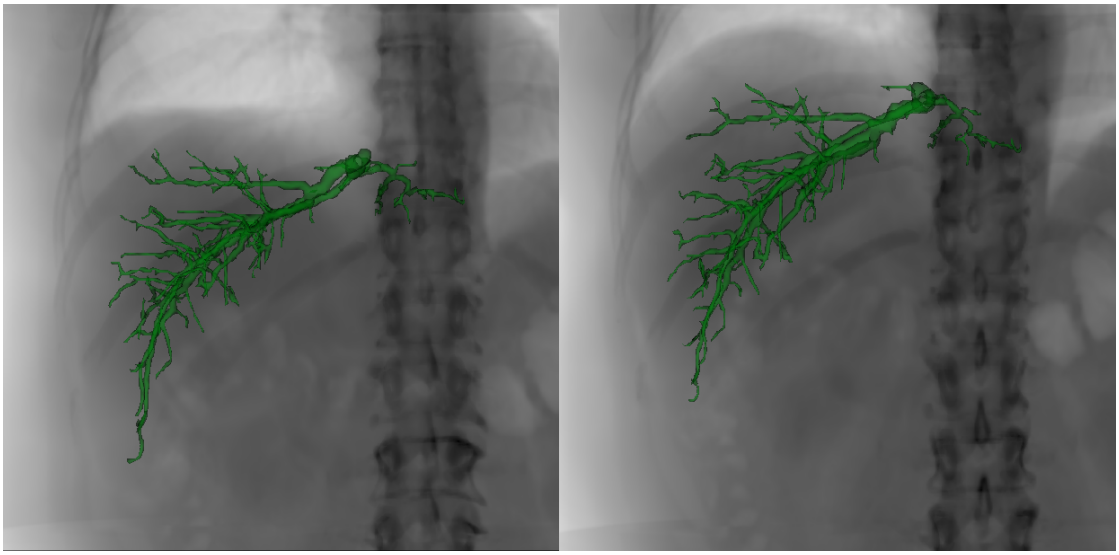


Figure 7.3: Images fluoroscopique synthétiques augmentées avec la position prédite des veines hépatiques lors de l'inspiration (gauche), et de l'expiration (droite)

Recalage des déformations interventionnelles

La troisième étude que nous avons réalisée, plus détaillée, a été soumise à publication dans le journal 'Medical Image Analysis'. Cette étude utilise la dernière version de notre méthode, présentée dans la section 7, et évalue la capacité de notre méthode à compenser les déformations respiratoires et arbitraires pouvant survenir lors d'une intervention. Nous utilisons une base de données expérimentale, acquise sur des modèles porcins avant et après intervention, ainsi que des données générées de façon synthétique pour mesurer les performances de notre méthode pour la prédiction de la position des veines hépatiques dans les images fluoroscopiques.

La base de données expérimentale, nommée IHUdeLiver10, est constituée de 10 paires de volumes CT pré- et post-intervention, acquises sur des modèles porcin. Nous utilisons quatre paires de volumes dans cette base de données, représentant les scénarios suivants : interactions d'une aiguille avec les tissus, mouvements anatomiques entre l'acquisition préopératoire et peropératoire, et chirurgie laparoscopique. Contrairement aux base de données disponibles dans la littérature, cette base de données contient des paires de volumes CT avant et après intervention, permettant d'évaluer la précision de méthodes conçues pour un usage interventionnel. Cependant, l'acquisition de volumes CT après injection d'agents de contraste ne produisant pas de résultats reproductibles, les arbres vasculaires diffèrent entre les deux volumes de chaque paire. Pour diminuer l'impact de ce problème et obtenir des arbres vasculaires comparables entre les données pré- et post-intervention,

Sujet n°	Dice 2D	
	Recalé	**Non recalé**
1	0,651	0,416
2	0,694	0,558
3	0,655	0,278
4	0,593	0,476

Table 7.1: Précision de notre méthode pour le recalage 2D-3D des veines hépatiques pour chaque sujet de la base de données de modèles porcins.

nous avons manuellement modifié les arbres vasculaires en enlevant les branches n'étant pas présentes dans les deux arbres. Pour chaque paire de cette base de données, nous avons généré une base de données d'entraînement à partir du volume CT pré-intervention, et une image fluoroscopique synthétique de test à partir du volume CT post-intervention. Après entraînement, nous avons mesuré le coefficient de Dice entre la projection en 2D des veines hépatiques prédites et des veines hépatiques réelles pour chaque volume CT de test, et reporté les résultats dans le tableau 7.1. En moyenne, notre méthode obtient un coefficient de Dice de $0,649 \pm 0,036$, avec un coefficient de Dice de $0,432 \pm 0,102$ avant recalage.

Pour évaluer notre méthode sur des données pour lesquelles la correspondance entre les arbres vasculaires avant et après intervention est parfaite, nous avons utilisé un volume CT pré-intervention acquis en routine clinique et nous avons appliqué des déformations synthétiques. La première base de données synthétiques vise à répliquer un mouvement respiratoire, comme dans l'étude précédente. Sur cette base de données, nous avons évalué les performances de notre réseau de neurones, entraîné sur des données agnostiques au domaine, ainsi que les performances de la méthode IGCN+ (Nakao *et al.*, 2022), développée pour la compensation du mouvement respiratoire. Sur les 50 images de test, représentant cinq périodes respiratoires, notre méthode obtient un coefficient de Dice moyen de $0,86 \pm 0,05$, tandis que la méthode IGCN+, spécifiquement entraînée sur le mouvement respiratoire, obtient un coefficient de Dice moyen de $0,88 \pm 0,04$ (coefficient de Dice moyen de $0,65 \pm 0,07$ avant recalage).

La seconde base de données synthétiques de test vise à répliquer une déformation induite par l'insertion d'une aiguille dans le foie. Sur les 50 images de test, notre méthode obtient un coefficient de Dice moyen de $0,80 \pm 0,01$, variant peu avec l'amplitude de déformation, tandis que la méthode IGCN+, entraîné sur le mouvement respiratoire, échoue à compenser cette déformation, obtenant un coefficient de Dice moyen de $0,70 \pm 0,02$, comparable au coefficient de Dice moyen de $0,69 \pm 0,03$ avant recalage. Nous avons également vérifié que notre méthode obte-

nait des performances similaires lorsqu'une aiguille était visible dans l'image, en ajoutant des aiguilles superposées aléatoirement aux images d'entraînement pour rendre le réseau de neurones robuste à la présence d'aiguilles dans l'image.

Afin de définir les meilleurs paramètres pour l'architecture du réseau et la fonction de coût, nous avons réalisé une analyse de sensibilité sur la base de données expérimentale IHUdeLiver10. Nous avons d'abord testé différentes fonction de coût :

- Erreur quadratique moyenne (EQM) entre le champ de déplacement prédit et de référence en 3D
- EQM entre le champ de déplacement prédit et de référence projetés en 2D (EQM 2D)
- EQM 2D + fonction de coût basée sur le Dice entre les segmentations 3D prédites et de référence.
- EQM 2D + fonction de coût basée sur le Dice entre les segmentations 2D prédites et de référence.

Parmi ces différentes fonctions de coût, la fonction de coût combinant l'EQM en 2D et le Dice en 2D (équation 7.3) donne les meilleurs résultats. Suite à ces résultats, nous avons testé différentes valeurs pour le paramètre de combinaison λ de la fonction de coût combinée, et nous avons obtenu les meilleurs résultats avec $\lambda = 0,5$. Nous avons également comparé les variantes de notre réseau avec et sans rétroprojection pour transformer les 'feature maps' 2D en 3D et obtenu les meilleurs résultats avec la rétroprojection. Finalement, nous avons aussi montré qu'utiliser l'augmentation de données (section 7) améliorait les performances de notre réseau, même si cette méthode a été développée pour pallier le changement d'apparence entre les images synthétiques et réelles.

Cette étude a permis de valider de façon approfondie l'utilité de notre méthode pour les interventions guidées par fluoroscopie, et sa superiorité dans ce cadre par rapport aux méthodes existantes. Grâce à l'analyse de sensibilité, nous avons pu améliorer les performances de notre méthode et valider les choix d'architecture et de fonction de coût.

Navigation endovasculaire autonome

La dernière étude de cette section concerne l'application de notre méthode à la navigation endovasculaire autonome. En utilisant notre méthode pour compenser les déformations induites par la respiration ou le mouvement cardiaque, nous avons pu améliorer le taux de succès de la méthode de navigation endovasculaire autonome

développée par V. Scarponi *et al.* de 24% à 93%. Cette étude a donné lieu à une présentation à la conférence internationale International Conference on Intelligent Robots and Systems 2024 (IROS).

> Valentina Scarponi, François Lecomte, *et al.* (Oct. 2024). "Autonomous Guidewire Navigation in Dynamic Environments". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*

Recalage déformable 2D-3D biomécanique

Les méthodes de recalage basées sur la physique permettent d'incorporer des connaissances a priori sur le comportement des organes pour pallier le manque d'information dans les images médicales. Notre objectif est ici d'incorporer un modèle biomécanique à notre méthode pour rendre les déformations prédites plus précises et réalistes. Nous avons exploré deux approches pour intégrer des contraintes physiques dans notre méthode de recalage 2D-3D. La première utilise la méthode des éléments finis pour générer des déformations physiquement plausibles lors de l'entraînement du réseau, améliorant ainsi sa précision sur des cas réels. Cette approche a été présentée dans le 'workshop' Data Curation and Augmentation in Enhancing Medical Imaging Applications (DCAMI) de la conférence internationale CVPR 2024.

> François Lecomte *et al.* (June 2024). "Beyond Respiratory Models: A Physics-enhanced Synthetic Data Generation Method for 2D-3D Deformable Registration". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2413–2421

La seconde approche vise à régulariser directement les prédictions du réseau pendant l'entraînement via des contraintes biomécaniques. Les expériences montrent qu'une régularisation en plusieurs étapes, utilisant d'abord l'élasticité linéaire puis l'hyperélasticité, permet d'obtenir des déformations plus réalistes tout en maintenant la précision du recalage.

Génération de données physiquement réalistes

Dans notre première approche, le champ de déformation généré aléatoirement est corrigé par un modèle biomécanique pour garantir le respect des lois physiques à l'intérieur du foie. Cette correction s'appuie sur un modèle hyperélastique néo-Hookéen résolu par éléments finis, avec des conditions aux limites de Dirichlet au

bords du foie garantissant la continuité avec le champ de déplacement aléatoire à l'extérieur du foie. Le champ final combine la solution physiquement réaliste à l'intérieur du foie avec le champ de déplacement original à l'extérieur, assurant des déformations réalistes pour les structures anatomiques d'intérêt comme les tumeurs et les vaisseaux.

Nous avons validé expérimentalement cette approche sur une paire de volumes CT issue de la base de données IHUdeLiver10 ainsi que sur des données synthétiques. Le volume CT synthétique consiste en un cube de taille similaire au foie du modèle porcin, avec un motif en damier d'intensités croissantes. Des déformations ont été appliquées en imposant des déplacements de -40 mm à +40 mm sur une face du cube, modélisé comme un solide hyperélastique de Mooney-Rivlin. Cette approche permet d'évaluer la précision de la méthode indépendamment des particularités anatomiques, tout en utilisant un modèle constitutif différent de celui employé pour l'entraînement.

Pour évaluer la précision du recalage, nous avons utilisé deux métriques différentes adaptées à chaque contexte : la distance de Wasserstein en 2D pour le modèle porcin où les correspondances point à point n'étaient pas disponibles, et la distance point à point moyenne pour le cube synthétique. La figure 7.4 montre les déformations de test pour le cube synthétique.

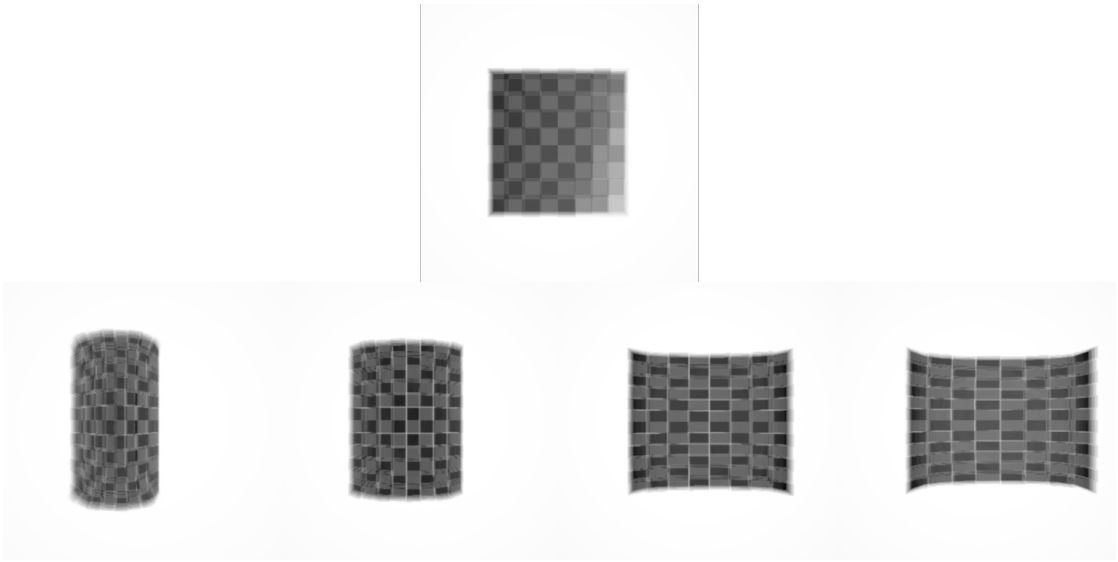


Figure 7.4: En haut, la fluoroscopie synthétique générée à partir du cube non déformé, et en bas, les fluoroscopies synthétiques générées à partir du cube déformé avec des déplacements de -40 mm, -20 mm, +20 mm et +40 mm (de gauche à droite).

Nos résultats montrent de meilleures performances pour les réseaux entraînés

avec la correction biomécanique du champ de déplacement sur les deux datasets, avec une erreur minimale de $2,8$ mm avec données régularisées contre $3,7$ mm sans régularisation (distance de Wasserstein) pour le modèle porcin. Pour le modèle synthétique, les gains de performances vont de $-4,6\%$ à $+50,1\%$, avec une moyenne de $+19,0\%$ (distance point à point). Ces résultats valident l'utilité d'introduire un modèle biomécanique pour corriger les déformations générées aléatoirement et ainsi obtenir de meilleures données d'entraînement pour le réseau.

Régularisation physique pendant l'entraînement

Pour notre seconde approche visant à incorporer les contraintes physiques pour obtenir des déformations prédites réalistes, nous avons exploré deux directions principales : l'amélioration de l'architecture du réseau pour permettre une régularisation physique efficace, et l'incorporation de la régularisation physique pendant l'entraînement.

Un régularisateur couramment utilisé pour imposer la régularité est le régularisateur de 'bending energy' $\mathcal{L}_{bending} = ||\nabla \mathbf{F}||^2$, avec $\mathbf{F}$ le gradient du champ de déplacement. Bien que produisant des déformations lisses, ce régularisateur tend à pénaliser l'amplitude des déformations. Pour pallier ce problème, nous avons utilisé un régularisateur basé sur la fonction de densité d'énergie de déformation hyperélastique d'un matériau néo-Hookéen. Ce régularisateur, $\mathcal{L}_{\nabla \cdot P}$, est basé sur la loi de conservation de la quantité de mouvement linéaire.

Pour ces deux régularisateurs, il est nécessaire de calculer les dérivées spatiales du champ de déformation prédit. Avec notre architecture d'origine, les dérivées spatiales ne peuvent pas être calculées en utilisant la différentiation automatique, mais doivent plutôt être approximées, par exemple en utilisant les différences finies.

Dans une première expérience, nous avons développé une variante nommée 'PointWiseDecoder' de notre architecture, utilisant des couches 'fully connected' pour rendre les prédictions différentiables par rapport aux coordonnées spatiales, en s'inspirant des méthodes de représentation neuronale implicite. Le réseau accepte en entrée des coordonnées spatiales et les 'feature maps' interpolées à ces positions pour prédire les déplacements correspondants. Bien que cette approche présente l'avantage théorique de pouvoir entraîner le réseau uniquement sur des points d'intérêt comme le maillage d'un organe, elle a obtenu des performances inférieures à notre architecture de base. Cette limitation peut s'expliquer par l'absence de mécanisme d'attention à longue portée, le réseau traitant les déplacements point à point.

Notre seconde approche a été d'utiliser une architecture de type HyperNetwork (Ha *et al.*, 2016) pour ajuster dynamiquement les paramètres des couches 'fully connected' décrites précédemment en fonction des 'feature maps', dans l'objectif de réintroduire une attention à longue portée. Comme nous n'avons pas pu

obtenir la convergence de cette variante lors de l'entraînement, nous avons conservé l'architecture convolutionelle initiale et utilisé les différences finies pour approximer les dérivées spatiales des déformations prédites.

Afin d'améliorer le réalisme des déformations prédites par le réseau de neurones, nous avons mené de nombreuses expériences de régularisation physique sur les jeux de données porcins et de cubes synthétique présentés ci-dessus. Nos expériences ont porté sur :

- Le test de plusieurs poids de régularisation pour équilibrer $\mathcal{L}_{\varphi^{2D}}$ et $\mathcal{L}_{\nabla \cdot P}$.
- L'application sélective de $\mathcal{L}_{\nabla \cdot P}$ en utilisant un masque où $\mathcal{L}_{\nabla \cdot P} \simeq 0$ dans le champ de déplacement de référence.
- La tentative de pré-entraînement avec uniquement $\mathcal{L}_{\varphi^{2D}}$ pour éviter la divergence précoce de $\mathcal{L}_{\nabla \cdot P}$ quand $J \to 0$.
- L'implémentation d'un modèle d'élasticité linéaire du foie pour la régularisation des déformations avec $\mathcal{L}^{lin}_{\nabla \cdot P}$.

Malgré ces différentes approches, nous n'avons pas réussi à obtenir une convergence simultanée de $\mathcal{L}_{\nabla \cdot P}$ et $\mathcal{L}_{\varphi^{2D}}$ pendant l'entraînement.

Pour vérifier qu'il était effectivement possible d'utiliser ce régularisateur physique pendant l'entraînement, nous avons mené des expériences sur un cas simplifié de recalage 2D. Dans ces expérience, l'objectif était de recaler une image binaire 2D représentant une coupe du foie ayant subi des déformations physiques (modèle néo-Hookéen). Nous avons utilisé une version simplifiée de notre architecture, opérant uniquement en 2D avec une structure encodeur-décodeur classique. Cette expérience a permis de valider l'approche de régularisation physique sur un cas simple avant d'envisager son application au cas plus complexe du recalage 2D-3D.

Dans notre première expérience, nous avons tenté d'entraîner directement le réseau en utilisant une combinaison de fonction de coût quadratique sur le champ de déplacement et de $\mathcal{L}_{\nabla \cdot P}$. Malgré l'utilisation d'un facteur d'échelle pour équilibrer les deux fonctions de coût, nous n'avons pas pu obtenir la convergence de l'entraînement. Cela est probablement dû au fait que la valeur de $\nabla \cdot P$ évolue de façon hautement non linéaire avec l'amplitude de déformations pour les déformations non physiques, contrairement à la fonction de coût quadratique, rendant difficile la minimisation conjointe.

Pour notre seconde expérience, nous avons développé une stratégie d'entraînement en plusieurs étapes pour incorporer des contraintes biomécaniques pendant l'entraînement du réseau. Cette approche consiste à entraîner d'abord le réseau avec une fonction de coût quadratique, puis à poursuivre l'entraînement en ajoutant $\mathcal{L}^{lin}_{\nabla \cdot P}$, puis finalement avec $\mathcal{L}_{\nabla \cdot P}$. Cette stratégie en plusieurs étapes permet d'obtenir la convergence de l'entraînement en combinant la fonction de coût sur le déplacement et

$\mathcal{L}_{\nabla \cdot P}$. Grâce à l'utilisation de la régularisation physique, le réalisme des déformations prédites est amélioré de façon significative, comme illustré par la figure 7.5.

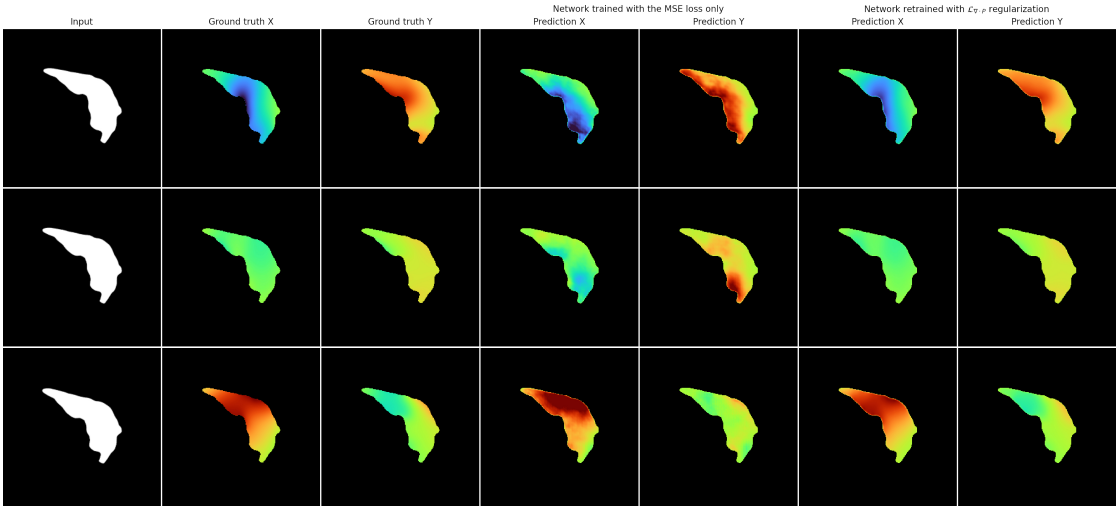


Figure 7.5: Comparaison qualitative entre les déformations prédites avec et sans $\mathcal{L}_{\nabla \cdot P}$.

Expériences sur images fluoroscopiques réelles

Pour valider notre méthode sur des images fluoroscopiques réelles, nous avons réalisé des expériences sur deux jeux de données. La première base de données, acquise sur un modèle porcin, contient un scan CT préopératoire avec 12 marqueurs radio-opaques implantés et une séquence de 74 images fluoroscopiques acquises pendant la respiration. La seconde base de données, acquise en routine clinique, contient un scan CT préopératoire, deux volumes CBCT peropératoires et les images fluoroscopiques associées.

Pour ces deux bases de données, nous avons d'abord estimé la pose des images fluoroscopiques par rapport au CT préopératoire en utilisant un processus en trois étapes :

1. Raffinement manuel de la pose avec DiffDRR, en ajustant itérativement les paramètres de translation et rotation jusqu'à obtenir une correspondance visuelle satisfaisante entre l'image fluoroscopique et le DRR.

2. Raffinement automatique initial avec DiffPose (Gopalakrishnan, Dey, *et al.*, 2024), en entraînant un réseau de neurones convolutif ResNet18 sur des don-

nées synthétiques générées avec des poses aléatoires autour de l'estimation initiale.

3. Raffinement final avec le module d'optimisation itérative de DiffPose, utilisant DiffDRR pour optimiser automatiquement les paramètres de pose en minimisant une fonction de coût basée sur l'indice de similarité structurelle (SSIM).

Ce processus en trois étapes a permis d'obtenir des DRR correspondant étroitement aux images fluoroscopiques pour le jeu de données porcin. Pour le jeu de données clinique, ce processus a échoué et nous avons utilisé la pose trouvée manuellement à l'étape 1. Ensuite, nous avons généré des données d'entraînement synthétiques à partir du CT préopératoire en utilisant les poses estimées.

Sur la base de données porcine, nous avons obtenu une erreur médiane de 2,4 mm en 2D sur la position des marqueurs radio-opaques. Cette erreur a été obtenue après plusieurs expériences visant à optimiser les paramètres d'entraînement du réseau : nous avons utilisé la procédure d'entraînement OneCycleLR pour accélerer la convergence de l'entraînement, nous avons résolu une erreur dans le module de rétroprojection, sans amélioration des performances, et nous avons étudié l'impact de la taille du batch sur les performances du réseau. Nos expériences ont notamment montré qu'un changement d'apparence des images dû à la présence d'intestins partiellement remplis pouvait être la cause l'échec du recalage pour un marqueur. Pour résoudre ce problème, nous avons essayé d'ajouter des variations d'intensité dans les images d'entraînement et ainsi améliorer la robustesse du réseau, mais cela n'a pas amélioré ses performances.

Sur la base de données clinique, les résultats sont plus difficiles à interpréter en raison de l'incertitude sur la pose des images fluoroscopiques et du changement de référentiel entre le CT préopératoire et les volumes CBCT. Pour les améliorer, nous avons employé le modèle pré-entraîné MedSAM (Ma *et al.*, 2024) pour augmenter l'image en entrée du réseau avec une segmentation 2D estimée du poumon, sans succès. Nous avons également essayé différents paramètres pour le réseau et des données d'entraînement avec plus ou moins de variations de pose, sans obtenir de performances satisfaisantes. Nous avons obtenu une erreur de 5,7 mm (5,1 mm avant recalage) en 3D sur la position d'un nodule pulmonaire, mais cette erreur est à relativiser car elle dépend fortement de la précision du recalage rigide 3D-3D entre le CT préopératoire et les volumes CBCT, et de l'estimation de pose des images fluoroscopiques.

Ces expériences ont permis de valider le potentiel de notre méthode pour les interventions guidées par fluoroscopie, tout en mettant en évidence certaines limitations, notamment dues au changement d'apparence dans les images ne résultant pas d'une déformation et à la difficulté d'estimer la pose des images fluoroscopiques

réelles. Pour poursuivre le développement de la méthode vers une utilisation clinique, il sera nécessaire de créer une base de données de test plus contrôlée, dans laquelle la pose des images fluoroscopiques serait connue avec précision.

Conclusion

L'adoption croissante des procédures mini-invasives a renforcé l'importance des solutions de navigation peropératoire dans la pratique clinique.

Notre méthode de recalage déformable 2D-3D s'intègre naturellement dans le flux de travail clinique existant, exploitant le scan CT préopératoire pour augmenter les images fluoroscopiques peropératoires. Notre approche de génération de données agnostique au domaine permet d'entraîner un réseau de neurones à retrouver des déformations arbitraires, dépassant ainsi les limitations des méthodes existantes focalisées sur le mouvement respiratoire.

Les études que nous avons menées sur données synthétiques et réelles ont permis de valider le potentiel de notre approche pour de futures applications cliniques. Dans une première étude, nous avons démontré la supériorité de notre méthode de génération de données agnostique au domaine par rapport à un modèle de déformation statistique. Les études suivantes ont confirmé son utilité clinique, notamment pour la visualisation des vaisseaux sans agents de contraste et la compensation des déformations liées aux interventions. L'intégration réussie avec un système de navigation endovasculaire autonome a également démontré son potentiel pratique.

L'incorporation de modèles biomécaniques a permis d'améliorer les performances de notre méthode en générant des données d'entraînement physiquement réalistes. De plus, les expériences sur un problème simplifié de recalage 2D ont mis en avant le potentiel de la régularisation biomécanique pour améliorer le réalisme des déformations prédites, soulignant le besoin de poursuivre les recherches dans cette voie.

Enfin, nous avons mené des expériences sur des données cliniques, montrant que notre processus d'entraînement sur des données synthétiques permettait au réseau de neurones de prédire la déformations dans des images fluoroscopiques réelles. Ces expériences ont également mis en évidence certaines limitations de notre approche, notamment la nécessité d'une base de données de test plus contrôlée pour poursuivre le développement de notre méthode.

Pour progresser vers une implémentation clinique, plusieurs axes de développement sont envisagés :

- L'amélioration de la robustesse aux variations anatomiques non liées aux déformations, via un processus de génération de données plus sophistiqué
- La modernisation de l'architecture du réseau, notamment par l'utilisation de

transformers et le développement de modèles fondamentaux pour le recalage 2D-3D

- L'intégration plus poussée des contraintes physiques dans l'architecture du réseau, suivant par exemple les principes des 'physics-augmented neural networks' (PANNs) (Linden *et al.*, 2023).

Bien que plusieurs directions de recherche restent à explorer pour une implémentation clinique complète, nos résultats actuels fournissent une base solide pour les développements futurs tout en confirmant le potentiel de notre approche.

Bibliography

Abi-Jaoudeh, Nadine *et al.* (2015). "Clinical experience with cone-beam CT navigation for tumor ablation". In: *Journal of Vascular and Interventional Radiology* 26.2, pp. 214–219.

Aboudara, Matt *et al.* (2020). "Improved diagnostic yield for lung nodules with digital tomosynthesis-corrected navigational bronchoscopy: initial experience with a novel adjunct". In: *Respirology* 25.2, pp. 206–213.

Adler, John R. *et al.* (1997). "The Cyberknife: A frameless robotic system for radiosurgery". In: *Stereotactic and Functional Neurosurgery*. Vol. 69. Issue: 1-4 ISSN: 10116125, pp. 124–128.

Al-Ahmad, Omar *et al.* (2020). "Improved FBG-Based Shape Sensing Methods for Vascular Catheterization Treatment". In: *IEEE Robotics and Automation Letters* 5.3, pp. 4687–4694.

Alkatout, Ibrahim *et al.* (2021). "The development of laparoscopy—a historical overview". In: *Frontiers in surgery* 8, p. 799442.

Allaire, Grégoire (2006). "Conception optimale de structures". In: *Conception optimale de structures*. Publisher: Springer Berlin Heidelberg.

Alvarez, Pablo, Matthieu Chabanas, *et al.* (2022). "Measurement and analysis of lobar lung deformation after a change of patient position during video-assisted thoracoscopic surgery". In: *IEEE Transactions on Biomedical Engineering* 70.3, pp. 931–940.

Alvarez, Pablo and Stéphane Cotin (2024). "Deformable Image Registration with Stochastically Regularized Biomechanical Equilibrium". In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 1–5.

Alvarez, Pablo, Simon Rouzé, *et al.* (Apr. 2021). "A hybrid, image-based and biomechanics-based registration approach to markerless intraoperative nodule localization during video-assisted thoracoscopic surgery". In: *Medical Image Analysis* 69, p. 101983.

Alvarez, Pablo A (2020). "Lung deformation estimation using a hybrid image-based/biomechanics-based approach for the localization of pulmonary nodules during video-assisted thoracoscopic surgery". PhD thesis. Université de Rennes.

Asano, F *et al.* (2002). "Virtual bronchoscopy in navigation of an ultrathin bronchoscope". In: *J Jpn Soc Bronchol* 24.6, pp. 433–8.

Asano, Fumihiro *et al.* (2006). "A virtual bronchoscopic navigation system for pulmonary peripheral lesions". In: *Chest* 130.2, pp. 559–566.

Avants, B. B. *et al.* (Feb. 2008). "Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain". In: *Medical image analysis* 12.1. Publisher: NIH Public Access, p. 26. ISSN: 13618415.

Avants, Brian B. *et al.* (Feb. 2011). "A Reproducible Evaluation of ANTs Similarity Metric Performance in Brain Image Registration". In: *NeuroImage* 54.3. Publisher: NIH Public Access, p. 2033. ISSN: 10538119.

Balakrishnan, Guha *et al.* (2019). "Voxelmorph: a learning framework for deformable medical image registration". In: *IEEE transactions on medical imaging* 38.8, pp. 1788–1800.

Baldwin, Andrew CW *et al.* (2016). "Through the looking glass: real-time video using 'Smart'technology provides enhanced intraoperative logistics". In: *World journal of surgery* 40, pp. 242–244.

Barnes, Connelly *et al.* (2009). "PatchMatch: A randomized correspondence algorithm for structural image editing". In: *ACM Trans. Graph.* 28.3, p. 24.

Barrow, Harry G *et al.* (1977). "Parametric correspondence and chamfer matching: Two new techniques for image matching". In: *Proceedings: Image Understanding Workshop*. Science Applications, Inc, pp. 21–27.

Bawaadam, Hasnain *et al.* (2024). "Integration of adjunct imaging for peripheral lung nodule sampling: a comprehensive review". In: *AME Medical Journal* 9.

Baydin, Atilim Gunes *et al.* (2018). "Automatic differentiation in machine learning: a survey". In: *Journal of machine learning research* 18.153, pp. 1–43.

Bellman, Richard (1957). "A Markovian Decision Process". In: *Journal of Mathematics and Mechanics* 6.5, pp. 679–684. ISSN: 00959057, 19435274.

Benameur, Said *et al.* (2003). "3D/2D registration and segmentation of scoliotic vertebrae using statistical models". In: *Computerized Medical Imaging and Graphics* 27.5, pp. 321–337.

Berger, Martin *et al.* (2016). "Marker-free motion correction in weight-bearing cone-beam CT of the knee joint". In: *Medical physics* 43.3, pp. 1235–1248.

Bisset, RAL, Ali N Khan, *et al.* (2012). *Differential diagnosis in abdominal ultrasound*. Elsevier India.

Bitar, Ibrahim *et al.* (2015). "A review on various formulations of displacement based multi-fiber straight Timoshenko beam finite elements". In: *Proc. CIGOS*.

Bonnans, J. Frédéric *et al.* (2003a). "Numerical Optimization: Theoretical and Practical Aspects". In: 1st ed. Springer. Chap. Newtonian Methods, pp. 51–66. ISBN: 3540001913; 9783540001911.

— (2003b). "Numerical Optimization: Theoretical and Practical Aspects". In: 1st ed. Springer. Chap. General Introduction, pp. 10–12. ISBN: 3540001913; 9783540001911.
Boussot, Valentin and Jean-Louis Dillenseger (2022). "Modèle statistique pour la prédiction de la déformation du poumon pendant la chirurgie thoracique vidéo-assistée". In: *RITS (Recherche en Imagerie et Technologies pour la Santé) 2022*.
— (2023). "Statistical model for the prediction of lung deformation during video-assisted thoracoscopic surgery". In: *Medical Imaging 2023: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 12466. SPIE, pp. 182–191.
Brock, KK *et al.* (2005). "Accuracy of finite element model-based multi-organ deformable image registration". In: *Medical physics* 32.6Part1, pp. 1647–1659.
Broit, Chaim (1981). *Optimal registration of deformed images*. University of Pennsylvania.
Brown, Mark A and Richard C Semelka (2011). *MRI: basic principles and applications*. John Wiley & Sons.
Buia, Alexander, Florian Stockhausen, and Ernst Hanisch (2015). "Laparoscopic surgery: a qualified systematic review". In: *World journal of methodology* 5.4, p. 238.
Cai, Shengze *et al.* (2021). "Physics-informed neural networks (PINNs) for fluid mechanics: A review". In: *Acta Mechanica Sinica* 37.12, pp. 1727–1738.
Campbell-Washburn, Adrienne E *et al.* (2017). "Real-time MRI guidance of cardiac interventions". In: *Journal of Magnetic Resonance Imaging* 46.4, pp. 935–950.
Celi, Simona *et al.* (2017). "Multimodality imaging for interventional cardiology". In: *Current pharmaceutical design* 23.22, pp. 3285–3300.
Chang, Chih-Wei *et al.* (2022). "A deep learning approach to transform two orthogonal X-ray images to volumetric images for image-guided proton therapy". In: *Medical Imaging 2022: Image Processing*. Vol. 12032. SPIE, pp. 484–490.
Chen, Junyu *et al.* (2022). "Transmorph: Transformer for unsupervised medical image registration". In: *Medical image analysis* 82, p. 102615.
Chen, Minheng, Zhirun Zhang, Shuheng Gu, Zhangyang Ge, *et al.* (2024). "Fully Differentiable Correlation-driven 2D/3D Registration for X-ray to CT Image Fusion". In: *arXiv preprint arXiv:2402.02498*.
Chen, Minheng, Zhirun Zhang, Shuheng Gu, and Youyong Kong (2024). "Embedded Feature Similarity Optimization with Specific Parameter Initialization for 2D/3D Medical Image Registration". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1521–1525.
Chen, Xiang *et al.* (2021). "Deep learning in medical image registration". In: *Progress in Biomedical Engineering* 3.1, p. 012003.
Chi, Wenqiang, Giulio Dagnino, *et al.* (2020). "Collaborative Robot-Assisted Endovascular Catheterization with Generative Adversarial Imitation Learning".

In: *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2414–2420.

Chi, Wenqiang, Jindong Liu, *et al.* (June 2018). "Learning-based endovascular navigation through the use of non-rigid registration for collaborative robotic catheterization". In: *International Journal of Computer Assisted Radiology and Surgery* 13 (6), pp. 855–864. ISSN: 18616429.

Choi, Jin Woo *et al.* (2012). "C-arm cone-beam CT–guided percutaneous transthoracic needle biopsy of small ($\leq$ 20 mm) lung nodules: diagnostic accuracy and complications in 161 patients". In: *American Journal of Roentgenology* 199.3, W322–W330.

Chou, Chen Rui and Stephen Pizer (2012). "Real-Time 2D/3D Deformable Registration Using Metric Learning". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7766 LNCS. Publisher: Springer, Berlin, Heidelberg ISBN: 9783642366192, pp. 1–10. ISSN: 03029743.

Chou, Chen-Rui, Brandon Frederick, *et al.* (2013). "2D/3D image registration using regression learning". In: *Computer Vision and Image Understanding* 117.9, pp. 1095–1106.

Chou, Chen-Rui and Stephen Pizer (2013). "Local Regression Learning via Forest Classification for 2D/3D Deformable Registration". In.

Cicenia, Joseph *et al.* (2021). "Augmented fluoroscopy: a new and novel navigation platform for peripheral bronchoscopy". In: *Journal of Bronchology & Interventional Pulmonology* 28.2, pp. 116–123.

Ciresan, Dan Claudiu *et al.* (2011). "Convolutional neural network committees for handwritten character classification". In: *2011 International conference on document analysis and recognition*. IEEE, pp. 1135–1139.

Cleary, Kevin and Terry M Peters (2010). "Image-guided interventions: technology review and clinical applications". In: *Annual review of biomedical engineering* 12.1, pp. 119–142.

Dahmen, Jessamyn and Diane Cook (2019). "SynSys: A synthetic data generation system for healthcare applications". In: *Sensors* 19.5, p. 1181.

Dataset, ADNI (n.d.). *ADNI Dataset*. `http://adni.loni.usc.edu/`.

Dataset, IXI (n.d.). *IXI Dataset*. `https://brain-development.org/ixi-dataset/`.

De Paolis, Lucio Tommaso and Valerio De Luca (2019). "Augmented visualization with depth perception cues to improve the surgeon's performance in minimally invasive surgery". In: *Medical & biological engineering & computing* 57, pp. 995–1013.

De Paolis, Lucio Tommaso and Francesco Ricciardi (2018). "Augmented visualisation in the treatment of the liver tumours with radiofrequency ablation". In:

Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 6.4, pp. 396–404.

De Vos, Bob D *et al.* (2017). "End-to-end unsupervised deformable image registration with a convolutional neural network". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, pp. 204–212.

Dea, Nicolas *et al.* (2016). "Economic evaluation comparing intraoperative cone beam CT-based navigation and conventional fluoroscopy for the placement of spinal pedicle screws: a patient-level data cost-effectiveness analysis". In: *The Spine Journal* 16.1, pp. 23–31. ISSN: 1529-9430.

Detmer, Felicitas J *et al.* (2017). "Virtual and augmented reality systems for renal interventions: A systematic review". In: *IEEE reviews in biomedical engineering* 10, pp. 78–94.

Doersch, Carl and Andrew Zisserman (2019). "Sim2real transfer learning for 3d human pose estimation: motion to the rescue". In: *Advances in Neural Information Processing Systems* 32.

Dosovitskiy, Alexey *et al.* (2020). "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929*.

Douglas, Bruce R, J William Charboneau, and Carl C Reading (2001). "Ultrasound-guided intervention: expanding horizons". In: *Radiologic Clinics of North America* 39.3, pp. 415–428.

Durrleman, Stanley *et al.* (Nov. 2014). "Morphometry of anatomical shape complexes with dense deformations and sparse parameters". In: *NeuroImage* 101, pp. 35–49. ISSN: 10538119.

El Hadramy, Sidaty, Nicolas Padoy, and Stéphane Cotin (2024). "HyperU-Mesh: Real-time deformation of soft-tissues across variable patient-specific parameters". In.

Epstein, Andrew J *et al.* (2013). "Impact of minimally invasive surgery on medical spending and employee absenteeism". In: *JAMA surgery* 148.7, pp. 641–647.

Falcoz, Pierre-Emmanuel *et al.* (Apr. 2015). " Video-assisted thoracoscopic surgery versus open lobectomy for primary non-small-cell lung cancer: a propensity-matched analysis of outcome from the European Society of Thoracic Surgeon database". In: *European Journal of Cardio-Thoracic Surgery* 49.2, pp. 602–609. ISSN: 1010-7940.

Faure, Francois *et al.* (June 2012). "SOFA: A Multi-Model Framework for Interactive Physical Simulation". In: *Soft Tissue Biomechanical Modeling for Computer Assisted Surgery*. Vol. 11. Springer, pp. 283–321. ISBN: 978-3-642-29013-8.

Fedorov, Andriy *et al.* (2012). "3D Slicer as an image computing platform for the Quantitative Imaging Network". In: *Magnetic resonance imaging* 30.9, pp. 1323–1341.

Felippa, Carlos A (2000). "A systematic approach to the element-independent corotational dynamics of finite elements". In.

Findl, Oliver *et al.* (2003). "Influence of operator experience on the performance of ultrasound biometry compared to optical biometry before cataract surgery". In: *Journal of Cataract & Refractive Surgery* 29.10, pp. 1950–1955.

Floridi, Chiara *et al.* (2017). "Clinical impact of cone beam computed tomography on iterative treatment planning during ultrasound-guided percutaneous ablation of liver malignancies". In: *Medical oncology* 34, pp. 1–8.

Foote, Markus *et al.* (2017). "Rank Constrained Diffeomorphic Density Motion Estimation for Respiratory Correlated Computed Tomography". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10551 LNCS. Publisher: Springer, Cham ISBN: 9783319676746, pp. 177–185. ISSN: 16113349.

Foote, Markus D. *et al.* (2019). "Real-Time 2D-3D Deformable Registration with Deep Learning and Application to Lung Radiotherapy Targeting". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11492 LNCS. Publisher: Springer Verlag, pp. 265–276.

Fu, Yabo *et al.* (Oct. 2020). "Deep learning in medical image registration: A review". In: *Physics in Medicine and Biology* 65.20. arXiv: 1912.12318 Publisher: IOP Publishing Ltd. ISSN: 13616560.

Gall, Kenneth P *et al.* (1993). "Experience using radiopaque fiducial points for patient alignment during radiotherapy". In: *International Journal of Radiation Oncology* Biology* Physics* 27, p. 161.

Gao, Cong, Anqi Feng, *et al.* (2023). "A fully differentiable framework for 2d/3d registration and the projective spatial transformers". In: *IEEE transactions on medical imaging*.

Gao, Cong, Benjamin D Killeen, *et al.* (2023). "Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis". In: *Nature Machine Intelligence* 5.3, pp. 294–308.

Gerard, Ian J *et al.* (2017). "Brain shift in neuronavigation of brain tumors: A review". In: *Medical image analysis* 35, pp. 403–420.

Ghibes, P. *et al.* (2023). "Endovascular treatment of symptomatic hepatic venous outflow obstruction post major liver resection". In: *BMC Gastroenterol* 23.1.

Ghosal, Sayan and Nilanjan Ray (2017). "Deep deformable registration: Enhancing accuracy by fully convolutional neural net". In: *Pattern Recognition Letters* 94, pp. 81–86.

Gislason-Lee, Amber J *et al.* (2016). "Impact of latest generation cardiac interventional X-ray equipment on patient image quality and radiation dose for trans-catheter aortic valve implantations". In: *The British journal of radiology* 89.1067, p. 20160269.

Glor, Fadi P *et al.* (2005). "Operator dependence of 3-D ultrasound-based computational fluid dynamics for the carotid bifurcation". In: *IEEE transactions on medical imaging* 24.4, pp. 451–456.

Golowa, Yosef and Jacob Cynamon (2012). "Endovascular Treatment of Portal Hypertension". In: *Haimovici's Vascular Surgery*. John Wiley & Sons, Ltd. Chap. 85, pp. 1095–1106. ISBN: 9781118481370.

Gopalakrishnan, Vivek, Neel Dey, and Polina Golland (2024). "Intraoperative 2d/3d image registration via differentiable x-ray rendering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11662–11672.

Gopalakrishnan, Vivek and Polina Golland (2022). "Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging". In: *Workshop on Clinical Image-Based Procedures*. Springer, pp. 1–11.

Gouveia, Ana R *et al.* (2012). "3D-2D image registration by nonlinear regression". In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 1343–1346.

Grimm, Matthias *et al.* (2021). "Pose-dependent weights and domain randomization for fully automatic X-ray to CT registration". In: *IEEE transactions on medical imaging* 40.9, pp. 2221–2232.

Guo, Jiaqi *et al.* (2024). "RN-SDEs: Limited-Angle CT Reconstruction with Residual Null-Space Diffusion Stochastic Differential Equations". In.

Ha, David, Andrew Dai, and Quoc V Le (2016). "Hypernetworks". In: *arXiv preprint arXiv:1609.09106*.

Haarnoja, Tuomas *et al.* (2018). *Soft Actor-Critic Algorithms and Applications*.

Hall, WA *et al.* (2003). *Costs and benefits of intraoperative MR-guided brain tumor resection*. Springer.

Hansen, Nikolaus, Sibylle D Müller, and Petros Koumoutsakos (2003). "Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES)". In: *Evolutionary computation* 11.1, pp. 1–18.

Harvey, CJ *et al.* (2012). "Applications of transrectal ultrasound in prostate cancer". In: *The British journal of radiology* 85.special_issue_1, S3–S17.

He, Fengxiang, Tongliang Liu, and Dacheng Tao (2019). "Control batch size and learning rate to generalize well: Theoretical and empirical evidence". In: *Advances in neural information processing systems* 32.

He, Fengxiang, Tongliang Liu, and Dacheng Tao (2020). "Why resnet works? residuals generalize". In: *IEEE transactions on neural networks and learning systems* 31.12, pp. 5349–5362.

He, Kaiming *et al.* (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hedstrom, Grady and Ajay A Wagh (2022). "Combining real-time 3-D imaging and augmented fluoroscopy with robotic bronchoscopy for the diagnosis of peripheral lung nodules". In: *Chest* 162.4, A2082.

Heizmann, Oleg *et al.* (2010). "Assessment of Intraoperative Liver Deformation During Hepatic Resection: Prospective Clinical Study". In: *World Journal of Surgery* 34, pp. 1887–1893.

Hirai, Ryusuke *et al.* (Mar. 2019). "Real-time tumor tracking using fluoroscopic imaging with deep neural network analysis". In: *Physica Medica* 59. Publisher: Associazione Italiana di Fisica Medica, pp. 22–29. ISSN: 1724191X.

Hoffmann, Malte *et al.* (2021). "SynthMorph: learning contrast-invariant registration without acquired images". In: *IEEE transactions on medical imaging* 41.3, pp. 543–558.

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks". In: *Neural networks* 4.2, pp. 251–257.

Hu, Xuemin *et al.* (2023). "How simulation helps autonomous driving: A survey of sim2real, digital twins, and parallel intelligence". In: *IEEE Transactions on Intelligent Vehicles*.

Huang, De-Xing *et al.* (2024). "Real-Time 2D/3D Registration via CNN Regression and Centroid Alignment". In: *IEEE Transactions on Automation Science and Engineering*.

Hugo, Geoffrey D. *et al.* (Feb. 2017). "A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer". In: *Medical physics* 44.2. Publisher: NIH Public Access, p. 762. ISSN: 00942405.

Iandola, Forrest *et al.* (2014). "Densenet: Implementing efficient convnet descriptor pyramids". In: *arXiv preprint arXiv:1404.1869*.

Ishida, Takashi *et al.* (2011). "Virtual bronchoscopic navigation combined with endobronchial ultrasound to diagnose small peripheral pulmonary lesions: a randomised trial". In: *Thorax* 66.12, pp. 1072–1077.

Izzo, Richard *et al.* (2018). "The vascular modeling toolkit: a python library for the analysis of tubular structures in medical images". In: *Journal of Open Source Software* 3.25, p. 745.

Jaganathan, Srikrishna *et al.* (2023). "Self-Supervised 2D/3D Registration for X-Ray to CT Image Fusion". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2788–2798.

Jourdan, Franck, Pierre Alart, and Michel Jean (1998). "A Gauss-Seidel like algorithm to solve frictional contact problems". In: *Computer Methods in Applied Mechanics and Engineering* 155.1, pp. 31–47. ISSN: 0045-7825.

Kandel, Ibrahem and Mauro Castelli (2020). "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset". In: *ICT express* 6.4, pp. 312–315.

Karstensen, Lennart, Tobias Behr, *et al.* (2020). "Autonomous guidewire navigation in a two dimensional vascular phantom". In: *Current Directions in Biomedical Engineering* 6.1.

Karstensen, Lennart, Jacqueline Ritter, *et al.* (Sept. 2023). "Recurrent neural networks for generalization towards the vessel geometry in autonomous endovascular guidewire navigation in the aortic arch". In: *Int. Journal of Computer Assisted Radiology and Surgery* 18 (9), pp. 1735–1744.

Ketelsen, Dominik *et al.* (2016). "Three-dimensional C-arm CT-guided transjugular intrahepatic portosystemic shunt placement: Feasibility, technical success and procedural time". In: *European radiology* 26, pp. 4277–4283.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kirillov, Alexander *et al.* (2023). "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026.

Kirk, Robert *et al.* (2023). "A Survey of Zero-shot Generalisation in Deep Reinforcement Learning". In: *Journal of Artificial Intelligence Research* 76, pp. 201–264.

Kitamura, Kei *et al.* (2002). "Registration accuracy and possible migration of internal fiducial gold marker implanted in prostate and liver treated with real-time tumor-tracking radiation therapy (RTRT)". In: *Radiotherapy and Oncology* 62.3, pp. 275–281. ISSN: 01678140.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2017). "ImageNet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6, pp. 84–90.

Kroes, Maarten W *et al.* (2016). "The use of laser guidance reduces fluoroscopy time for C-arm cone-beam computed tomography-guided biopsies". In: *Cardiovascular and Interventional Radiology* 39, pp. 1322–1326.

Kumar, Ajay and Alwin Chuan (2009). "Ultrasound guided vascular access: efficacy and safety". In: *Best Practice & Research Clinical Anaesthesiology* 23.3, pp. 299–311.

Kuzhagaliyev, Timur *et al.* (2018). "Augmented reality needle ablation guidance tool for irreversible electroporation in the pancreas". In: *Medical imaging 2018: Image-guided procedures, robotic interventions, and modeling*. Vol. 10576. SPIE, pp. 260–265.

Kweon, Jihoon *et al.* (2021). "Deep Reinforcement Learning for Guidewire Navigation in Coronary Artery Phantom". In: *IEEE Access* 9, pp. 166409–166422.

Lanfranchi, Filippo *et al.* (2024). "Use of the Archimedes navigation system to diagnose peripheral pulmonary lesions: preliminary Italian results". In: *Frontiers in Oncology* 14.

Lecomte, Francois *et al.* (2023). "Enhancing fluoroscopy-guided interventions: a neural network to predict vessel deformation without contrast agents". In: *The Hamlyn Symposium on Medical Robotics*. The Hamlyn Centre, Imperial College London London, UK, pp. 75–76.

Lecomte, François *et al.* (June 2024). "Beyond Respiratory Models: A Physics-enhanced Synthetic Data Generation Method for 2D-3D Deformable Registration". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2413–2421.

Lee, Brian C *et al.* (2022). "Breathing-Compensated Neural Networks for Real Time C-Arm Pose Estimation in Lung CT-Fluoroscopy Registration". In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 1–5.

Lee, Min Woo *et al.* (2010). "Targeted sonography for small hepatocellular carcinoma discovered by CT or MRI: factors affecting sonographic detection". In: *American Journal of Roentgenology* 194.5, W396–W400.

Lei, Yang, Zhen Tian, Tonghe Wang, Marian Axente, *et al.* (2022). "Fast 3D imaging via deep learning for deep inspiration breath-hold lung radiotherapy". In: *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 12034. SPIE, pp. 628–633.

Lei, Yang, Zhen Tian, Tonghe Wang, Justin Roper, *et al.* (2021). "Deep learning-based 3D image generation using a single 2D projection image". In: *Medical Imaging 2021: Image Processing*. Vol. 11596. SPIE, pp. 516–521.

Lesage, Anne-Cécile *et al.* (2020). "Preliminary evaluation of biomechanical modeling of lung deflation during minimally invasive surgery using pneumothorax computed tomography scans". In: *Physics in Medicine & Biology* 65.22, p. 225010.

Li, Lei and Suping Wu (2021). "Dmifnet: 3d shape reconstruction based on dynamic multi-branch information fusion". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 7219–7225.

Linden, Lennart *et al.* (2023). "Neural networks meet hyperelasticity: A guide to enforcing physics". In: *Journal of the Mechanics and Physics of Solids* 179, p. 105363.

Liu, David *et al.* (2010). "Monitoring with head-mounted displays in general anesthesia: a clinical evaluation in the operating room". In: *Anesthesia & Analgesia* 110.4, pp. 1032–1038.

Liu, Dong C and Jorge Nocedal (1989). "On the limited memory BFGS method for large scale optimization". In: *Mathematical programming* 45.1, pp. 503–528.

Loshchilov, Ilya and Frank Hutter (2017). "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101*.

— (2016). "Sgdr: Stochastic gradient descent with warm restarts". In: *arXiv preprint arXiv:1608.03983*.

Ma, Jun *et al.* (2024). "Segment Anything in Medical Images". In: *Nature Communications* 15, p. 654.

Maier, IL *et al.* (2018). "Diagnosing early ischemic changes with the latest-generation flat detector CT: a comparative study with multidetector CT". In: *American Journal of Neuroradiology* 39.5, pp. 881–886.

Mamoulakis, Charalampos *et al.* (2017). "Contrast-induced nephropathy: Basic concepts, pathophysiological implications and prevention strategies". In: *Pharmacology & therapeutics* 180, pp. 99–112.

Manhire, A *et al.* (2003). "Guidelines for radiologically guided lung biopsy". In: *Thorax* 58.11, pp. 920–936.

Mantz, J-M *et al.* (1982). "Le choc anaphylactique: Résultats d'une enquête nationale portant sur 1047 cas". In: *La Revue de Médecine Interne* 3.4, pp. 331–338.

Maybody, Majid, Carsten Stevenson, and Stephen B Solomon (2013). "Overview of navigation systems in image-guided interventions". In: *Techniques in vascular and interventional radiology* 16.3, pp. 136–143.

McClennan, Bruce L (1990). "Preston M. Hickey memorial lecture. Ionic and nonionic iodinated contrast media: evolution and strategies for use." In: *AJR. American journal of roentgenology* 155.2, pp. 225–233.

McCollough, Cynthia H *et al.* (2015). "Answers to common questions about the use and safety of CT scans". In: *Mayo Clinic Proceedings*. Vol. 90. 10. Elsevier, pp. 1380–1392.

Mert, Ayguel *et al.* (2012). "Brain tumor surgery with 3-dimensional surface navigation". In: *Operative Neurosurgery* 71, ons286–ons295.

Mhaskar, Hrushikesh N and Charles A Micchelli (1992). "Approximation by superposition of sigmoidal and radial basis functions". In: *Advances in Applied mathematics* 13.3, pp. 350–373.

Miao, Shun *et al.* (2018). "Dilated FCN for multi-agent 2D/3D medical image registration". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1.

Milletarì, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi (2016). "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571.

Miranda, Victor *et al.* (2023). "Generalization in Deep Reinforcement Learning for Robotic Navigation by Reward Shaping". In: *IEEE Transactions on Industrial Electronics*, pp. 1–8.

Modersitzki, Jan (2003). *Numerical methods for image registration*. OUP Oxford.

Mooney, Melvin (1940). "A theory of large elastic deformation". In: *Journal of applied physics* 11.9, pp. 582–592.

Nakao, Megumi, Mitsuhiro Nakamura, and Tetsuya Matsuda (2022). "Image-to-Graph Convolutional Network for 2D/3D Deformable Model Registration of Low-Contrast Organs". In: *IEEE Transactions on Medical Imaging* 41.12, pp. 3747–3761.

Noble, J Alison, Nassir Navab, and H Becher (2011). "Ultrasonic image analysis and image-guided interventions". In: *Interface focus* 1.4, pp. 673–685.

Ormerod, DF, B Ross, and A Naluai-Cecchini (2003). "Use of an augmented reality display of patient monitoring data to enhance anesthesiologists' response to abnormal clinical events". In: *Medicine Meets Virtual Reality 11*. IOS Press, pp. 248–250.

Ortega, G *et al.* (2008). "Usefulness of a head mounted monitor device for viewing intraoperative fluoroscopy during orthopaedic procedures". In: *Archives of orthopaedic and trauma surgery* 128, pp. 1123–1126.

Park, Brian J *et al.* (2020). "3D Augmented reality-assisted CT-Guided interventions: system design and preclinical trial on an abdominal phantom using HoloLens 2". In: *arXiv preprint arXiv:2005.09146*.

Paszke, Adam *et al.* (2017). "Automatic differentiation in pytorch". In.

Pennec, Xavier *et al.* (2005). "Riemannian elasticity: A statistical regularization framework for non-linear registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 943–950.

Pertzov, Barak *et al.* (2021). "The LungVision navigational platform for peripheral lung nodule biopsy and the added value of cryobiopsy". In: *Thoracic Cancer* 12.13, pp. 2007–2012.

Powell, Michael JD *et al.* (2009). "The BOBYQA algorithm for bound constrained optimization without derivatives". In: *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge* 26, pp. 26–46.

Pritchett, Michael A (2021). "Prospective analysis of a novel endobronchial augmented fluoroscopic navigation system for diagnosis of peripheral pulmonary le-

sions". In: *Journal of Bronchology & Interventional Pulmonology* 28.2, pp. 107–115.

Przkora, Rene *et al.* (2015). "Evaluation of the head-mounted display for ultrasound-guided peripheral nerve blocks in simulated regional anesthesia". In: *Pain Medicine* 16.11, pp. 2192–2194.

Puijk, Robbert S *et al.* (2018). "Percutaneous liver tumour ablation: image guidance, endpoint assessment, and quality control". In: *Canadian Association of Radiologists Journal* 69.1, pp. 51–62.

Puschel, A, C Schafmayer, and J Groß (2022). "Robot-assisted techniques in vascular and endovascular surgery". In: *Langenbecks Arch Surg* 407.5, pp. 1789–1795.

Rabbitt, Richard D *et al.* (1995). "Mapping of hyperelastic deformable templates using the finite element method". In: *Vision Geometry IV*. Vol. 2573. SPIE, pp. 252–265.

Radiuk, Pavlo M (2017). "Impact of training set batch size on the performance of convolutional neural networks for diverse datasets". In.

Raffin, Antonin *et al.* (2021). "Stable-Baselines3: Reliable Reinforcement Learning Implementations". In: *Journal of Machine Learning Research* 22.268, pp. 1–8.

Ranzato, Marc'Aurelio *et al.* (2006). "Efficient learning of sparse representations with an energy-based model". In: *Advances in neural information processing systems* 19.

Rehani, MM *et al.* (2010). "Radiological protection in fluoroscopically guided procedures performed outside the imaging department". In: *Annals of the ICRP* 40.6, pp. 1–102.

Rivlin, Ronald S (1948). "Large elastic deformations of isotropic materials IV. Further developments of the general theory". In: *Philosophical transactions of the royal society of London. Series A, Mathematical and physical sciences* 241.835, pp. 379–397.

Rouze, Simon *et al.* (2016). "Small pulmonary nodule localization with cone beam computed tomography during video-assisted thoracic surgery: a feasibility study". In: *Interactive CardioVascular and Thoracic Surgery* 22.6, pp. 705–711.

Rouzé, Simon (2022). "Localisation de nodules pulmonaires en chirurgie mini-invasive assistée par ordinateur". PhD thesis. Université de Rennes 1, pp. 62–63.

Sailer, Anna M *et al.* (2015). "Radiation exposure of abdominal cone beam computed tomography". In: *Cardiovascular and interventional radiology* 38, pp. 112–120.

Salvi, Andrey *et al.* (2020). "Attention-based 3D object reconstruction from a single image". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8.

Sarti, Marc, William P Brehmer, and Spencer B Gay (2012). "Low-dose techniques in CT-guided interventions". In: *Radiographics* 32.4, pp. 1109–1119.

Scarponi, Valentina, Michel Duprez, *et al.* (2024). "A zero-shot reinforcement learning strategy for autonomous guidewire navigation". In: *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–8.

Scarponi, Valentina, François Lecomte, *et al.* (Oct. 2024). "Autonomous Guidewire Navigation in Dynamic Environments". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Scarselli, Franco *et al.* (2008). "The graph neural network model". In: *IEEE transactions on neural networks* 20.1, pp. 61–80.

Schafer, Sebastian and Jeffrey H Siewerdsen (2020). "Technology and applications in interventional imaging: 2D X-ray radiography/fluoroscopy and 3D cone-beam CT". In: *Handbook of medical image computing and computer assisted intervention*. Elsevier, pp. 625–671.

Seppenwoolde, Yvette *et al.* (2011). "Treatment precision of image-guided liver SBRT using implanted fiducial markers depends on marker-tumour distance". In: *Article in Physics in Medicine and Biology*.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shan, Siyuan *et al.* (2017). "Unsupervised end-to-end learning for deformable medical image registration". In: *arXiv preprint arXiv:1711.08608*.

Shao, Hua-Chieh, Yunxiang Li, *et al.* (2023). "Real-time liver motion estimation via deep learning-based angle-agnostic X-ray imaging". In: *Medical physics* 50.11, pp. 6649–6662.

Shao, Hua-Chieh, Jing Wang, *et al.* (2022). "Real-time liver tumor localization via a single x-ray projection using deep graph neural network-assisted biomechanical modeling". In: *Physics in Medicine & Biology* 67.11, p. 115009.

Shen, Liyue, Wei Zhao, and Lei Xing (2019). "Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning". In: *Nature Biomedical Engineering* 3.11. Publisher: Springer US ISBN: 4155101904, pp. 880–888. ISSN: 2157846X.

Shi, Pengcheng *et al.* (2024). "Centerline Boundary Dice Loss for Vascular Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 46–56.

Shieh, Chun Chien *et al.* (Mar. 2017). "A Bayesian approach for three-dimensional markerless tumor tracking using kV imaging during lung radiotherapy". In: *Physics in Medicine and Biology* 62.8. Publisher: Institute of Physics Publishing, pp. 3065–3080. ISSN: 13616560.

Simonovsky, Martin *et al.* (2016). "A deep metric for multimodal registration". In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016:*

19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III 19. Springer, pp. 10–18.

Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Singh, S. *et al.* (2017). "Anatomic study of the morphology of the right and left coronary arteries". In: *Folia Morphologica* 76.4, pp. 668–674. ISSN: 1644-3284.

Smith, Leslie N and Nicholay Topin (2019). "Super-convergence: Very fast training of neural networks using large learning rates". In: *Artificial intelligence and machine learning for multi-domain operations applications*. Vol. 11006. SPIE, pp. 369–386.

Solbiati, Marco *et al.* (2018). "Augmented reality for interventional oncology: proof-of-concept study of a novel high-end guidance system platform". In: *European radiology experimental* 2, pp. 1–9.

Sotiras, Aristeidis, Christos Davatzikos, and Nikos Paragios (2010). "Deformable Medical Image Registration: A Survey". In: *IEEE transactions on medical imaging*.

Sun, Jiayuan *et al.* (2022). "Efficacy and safety of virtual bronchoscopic navigation with fused fluoroscopy and vessel mapping for access of pulmonary lesions". In: *Respirology* 27.5, pp. 357–365.

Szegedy, Christian *et al.* (2017). "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1.

Tang, Thomas SY, Randy E Ellis, and Gabor Fichtinger (2000). "Fiducial registration from a single X-Ray image: a new technique for fluoroscopic guidance and radiotherapy". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2000: Third International Conference, Pittsburgh, PA, USA, October 11-14, 2000. Proceedings 3*. Springer, pp. 502–511.

Tang, Yucheng *et al.* (2022). "Self-supervised pre-training of swin transformers for 3d medical image analysis". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20730–20740.

Thirion, J-P (1998). "Image matching as a diffusion process: an analogy with Maxwell's demons". In: *Medical image analysis* 2.3, pp. 243–260.

Tian, Lin *et al.* (2022). "LiftReg: Limited Angle 2D/3D Deformable Registration". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*. Springer, pp. 207–216.

Tian, Wei *et al.* (2023). "A DDPG-Based Method of Autonomous Catheter Navigation in Virtual Environment". In: *Proc. International Conference on Mechatronics and Automation*, pp. 889–893.

Tobin, Josh *et al.* (2017). "Domain randomization for transferring deep neural networks from simulation to the real world". In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, pp. 23–30.

Tonutti, Michele *et al.* (2017). "The role of technology in minimally invasive surgery: state of the art, recent developments and future directions". In: *Postgraduate medical journal* 93.1097, pp. 159–167.

Trouve, Alain *et al.* (2005). "Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms". In: *International Journal of Computer Vision* 61.2, pp. 139–157.

Unberath, Mathias *et al.* (Sept. 2018). "DeepDRR – A Catalyst for Machine Learning in Fluoroscopy-Guided Procedures". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11073 LNCS, pp. 98–106. ISSN: 16113349.

Van den Elsen, Petra A, E-JD Pol, and Max A Viergever (1993). "Medical image matching-a review with classification". In: *IEEE Engineering in Medicine and Biology Magazine* 12.1, pp. 26–39.

Varro, Zoltan, Julia K Locklin, and Bradford J Wood (2004). "Laser navigation for radiofrequency ablation". In: *Cardiovascular and interventional radiology* 27, pp. 512–515.

Vles, MD *et al.* (2020). "Virtual and augmented reality for preoperative planning in plastic surgical procedures: a systematic review". In: *Journal of Plastic, Reconstructive & Aesthetic Surgery* 73.11, pp. 1951–1959.

Wagh, A, E Ho, and K Hogarth (2021). "Combining the use of robotic bronchoscopy with augmented fluoroscopy to diagnose peripheral pulmonary lesions". In: *Clin Oncol* 6, p. 1811.

Wallace, Michael J *et al.* (2008). "Three-dimensional C-arm cone-beam CT: applications in the interventional suite". In: *Journal of Vascular and Interventional Radiology* 19.6, pp. 799–813.

Wang, Nanyang *et al.* (2018). "Pixel2mesh: Generating 3d mesh models from single rgb images". In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 52–67.

Wang, Shuang *et al.* (2022). "Study on Autonomous Delivery of Guidewire Based on Improved YOLOV5s on Vascular Model Platform". In: *2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1–6.

Wasserthal, Jakob *et al.* (2023). "TotalSegmentator: robust segmentation of 104 anatomic structures in CT images". In: *Radiology: Artificial Intelligence* 5.5.

Wei, Ran *et al.* (2020). "Real-time tumor localization with single x-ray projection at arbitrary gantry angles using a convolutional neural network (CNN)". In: *Physics in Medicine & Biology* 65.6, p. 065012.

Wijesinghe, W Okandapola Kankanamalage Isuru Suranga (2024). "Intelligent image-driven motion modelling for adaptive radiotherapy".
Winter, Jeff D *et al.* (2015). "Accuracy of robotic radiosurgical liver treatment throughout the respiratory cycle". In: *International Journal of Radiation Oncology* Biology* Physics* 93.4, pp. 916–924.
Wolterink, Jelmer M, Jesse C Zwienenberg, and Christoph Brune (2022). "Implicit neural representations for deformable image registration". In: *International Conference on Medical Imaging with Deep Learning*. PMLR, pp. 1349–1359.
Wu, Guorong *et al.* (2015). "Scalable high-performance image registration framework by unsupervised deep feature representations learning". In: *IEEE transactions on biomedical engineering* 63.7, pp. 1505–1516.
Wu, Yi-Wei *et al.* (2016). "Prevention and management of adverse reactions induced by iodinated contrast media". In: *Ann Acad Med Singapore* 45.4, pp. 157–164.
Wuerfel, Jens *et al.* (2004). "Changes in cerebral perfusion precede plaque formation in multiple sclerosis: a longitudinal perfusion MRI study". In: *Brain* 127.1, pp. 111–119.
Wunsch, Patrick and Gerhard Hirzinger (1996). "Registration of CAD-models to images by iterative inverse perspective matching". In: *Proceedings of 13th International Conference on Pattern Recognition*. Vol. 1. IEEE, pp. 78–83.
Yan, Yongxuan *et al.* (2024). "Markerless Lung Tumor Localization From Intraoperative Stereo Color Fluoroscopic Images for Radiotherapy". In: *IEEE Access* 12, pp. 40809–40826.
Yanovsky, Igor *et al.* (2008). "Unbiased volumetric registration via nonlinear elastic regularization". In: *2nd MICCAI workshop on mathematical foundations of computational anatomy*.
Yin, Wei *et al.* (2022). "Towards accurate reconstruction of 3d scene shape from a single monocular image". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.5, pp. 6480–6494.
Yoon, Jang W *et al.* (2018). "Augmented reality for the surgeon: systematic review". In: *The international journal of medical robotics and computer assisted surgery* 14.4, e1914.
You, Zhonghui *et al.* (2019). "Adversarial Noise Layer: Regularize Neural Network by Adding Noise; Adversarial Noise Layer: Regularize Neural Network by Adding Noise". In: *2019 IEEE International Conference on Image Processing (ICIP)*. ISBN: 9781538662496.
Zhang, Jingyu *et al.* (2024). "AI co-pilot bronchoscope robot". In: *Nature communications* 15.1, p. 241.

Zhang, Quncheng *et al.* (2021). "Combination of the Archimedes Navigation System and cryobiopsy in diagnosis of diffuse lung disease". In: *Journal of International Medical Research* 49.7, p. 03000605211016665.

Zhang, Y, X Huang, and J Wang (2020). "Automatic Cone Beam Projection-Based Liver Tumor Localization by Deep Learning and Biomechanical Modeling". In: *International Journal of Radiation Oncology, Biology, Physics* 108.3, S171.

Zhang, Y, X Huang, J Wang, *et al.* (Nov. 2020). "Automatic Cone Beam Projection-based Liver Tumor Localization by Deep Learning and Biomechanical Modeling". In: *International Journal of Radiation Oncology, Biology, Physics* 108.3. Publisher: Elsevier, S171. ISSN: 0360-3016.

Zhang, Yikun *et al.* (2021). "CLEAR: comprehensive learning enabled adversarial reconstruction for subtle structure enhanced low-dose CT imaging". In: *IEEE Transactions on Medical Imaging* 40.11, pp. 3089–3101.

Zhang, You (2021). "An unsupervised 2D–3D deformable registration network (2D3D-RegNet) for cone-beam CT estimation". In: *Physics in Medicine & Biology* 66.7, p. 074001.

Zhao, Wei *et al.* (2019). "Markerless pancreatic tumor target localization enabled by deep learning HHS Public Access". In: *Int J Radiat Oncol Biol Phys* 105.2, pp. 432–439.

Zhao, Zhuo *et al.* (2021). "Augmented reality technology in image-guided therapy: State-of-the-art review". In: *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* 235.12, pp. 1386–1398.

Zhu, Jun-Yan *et al.* (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.

Zhu, Yongpei and Shi Lu (2022). "Swin-voxelmorph: A symmetric unsupervised learning model for deformable medical image registration using swin transformer". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 78–87.

Zhuang, Xiahai and Yipeng Hu (2017). "Statistical Deformation Model: Theory and Methods". In: *Statistical Shape and Deformation Analysis*. Elsevier, pp. 33–65.

École doctorale **Mathématiques, sciences** de l'**information** et de l'**ingénieur** | ED 269
Université de Strasbourg

Inria

Titre: Amélioration des Procédures Guidées par Fluoroscopie à l'aide d'un Réseau de Neurones pour le Recalage Déformable des Organes

Mots clés: Recalage déformable, recalage 2D-3D, fluoroscopie, Réseaux de Neurones Convolutifs, génération de données synthétiques

Résumé: Dans les interventions guidées par fluoroscopie, le manque de contraste empêche la visualisation directe des structures anatomiques essentielles. Les solutions existantes présentent des inconvénients significatifs: l'utilisation de CBCT augmente l'exposition aux radiations, tandis que les agents de contraste présentent des risques de toxicité pour les patients. Les techniques de recalage fluoroscopie-CT pourraient résoudre ces problèmes, mais la littérature existante s'est principalement concentrée sur la compensation du mouvement respiratoire. Or, pendant les interventions, l'action des cliniciens sur les organes est également source de déformations, rendant ces approches de recalage inefficaces. Pour répondre à ces défis, nous présentons une méthode de recalage déformable 2D-3D en temps réel adaptée aux interventions guidées par fluoroscopie. Notre approche par apprentissage profond s'intègre dans la pratique clinique courante, avec un temps d'entraînement minimal après l'acquisition du scanner préopératoire. Grâce à notre processus de génération de données agnostique, le réseau de neurones entraîné est capable de compenser des déformations arbitraires, en exploitant les informations de pose avec son module de rétroprojection 2D-3D. Les expériences sur des images fluoroscopiques simulées ont montré la capacité de notre méthode à apporter une visualisation en temps réel des vaisseaux sans agents de contraste. Sur des images fluoroscopiques réelles, notre méthode permet de compenser le mouvement respiratoire avec une précision médiane de 2,4 mm. Ces résultats démontrent le potentiel de la méthode proposée, établissant une base pour de futurs développements tout en motivant la conduite d'une validation clinique plus aboutie.

Title: Enhancing Fluoroscopy-Guided Procedures with Neural Network-Based Deformable Organ Registration

Keywords: Deformable registration, 2D-3D registration, fluoroscopy, Convolutional Neural Networks, Synthetic data generation

Abstract: In fluoroscopy-guided interventions, the lack of contrast prevents direct visualization of essential anatomical structures. Existing solutions have significant drawbacks: the use of CBCT increases radiation exposure, while contrast agents present toxicity risks for patients. Fluoroscopy to CT registration has the potential to alleviate these issues, but existing literature has primarily focused on respiratory motion compensation. Yet, during interventions, clinicians' actions on organs are an additional source of deformation, rendering these registration approaches ineffective. To address these challenges, we present a real-time 2D-3D deformable registration method tailored to fluoroscopy-guided interventions. Our proposed deep learning approach seamlessly integrates into existing clinical workflows, with minimal training time after preoperative CT scan acquisition. Thanks to our novel domain-agnostic data generation framework, the trained neural network can recover arbitrary deformations, leveraging pose information through its 2D-3D feature backprojection module. Experiments on simulated fluoroscopic images demonstrated our method's ability to provide real-time vessel visualization without contrast agents. On real fluoroscopic images, our method compensates for respiratory motion with a median accuracy of 2.4 mm. These results demonstrate the potential of the proposed method, establishing a foundation for future developments while motivating more comprehensive clinical validation.