UNIVERSITY OF LEEDS

Intelligent Image-driven Motion Modelling for Adaptive Radiotherapy

W.O.K. Isuru Suranga Wijesinghe

University of Leeds

School of Mechanical Engineering

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

May 2024

This thesis is dedicated to:

every brave heart touched by the shadows of cancer

Intellectual Property Statement

The candidate confirms that the work submitted is his own except where work which has formed part of jointly authored publications has been included. The candidate confirms that appropriate credit has been given where reference has been made to the work of others. The contribution of the candidate and the other authors to this work has been explicitly indicated as follows.

Signed 

Acknowledgements

First and foremost, I am extremely grateful to my primary supervisor Dr Zeike Taylor (University of Leeds), and Dr Michael Nix (Leeds NHS Trust), my clinical supervisor, for their invaluable guidance, encouragement, patience, and dedicated involvement throughout my doctoral research. I will always appreciate the knowledge they shared with me and the opportunities they afforded. They have been a huge inspiration to me throughout this journey because of their incredible passion and determination, which has pushed me to always polish my critical thinking skills.

I would like to extend my deepest gratitude to Dr Ali Gooya (University of Glasgow), my secondary supervisor, for his immense support, valuable feedback, and guidance throughout the journey. I am also very thankful to Dr Arezoo Zakari, a research fellow, at the University of Leeds for providing me with enormous support both technically and non-technically.

My humble gratitude goes to the School of Mechanical Engineering at the University of Leeds for awarding me a fully funded scholarship, which provided great financial support throughout these four years. I would also like to acknowledge the immense support received from the RADNET group in Leeds.

I am extremely grateful to my colleagues at the University of Leeds who have encouraged me along this journey. Last but not least, I would like to express my deepest gratitude to my loving wife, Chathurika, and our respective families for being with me at every step in all my career choices, including this PhD.

Abstract

Internal anatomical motion (e.g. respiration-induced motion) confounds the precise delivery of radiation to target volumes during external beam radiotherapy. Precision is, however, critical to ensure prescribed radiation doses are delivered to the target (tumour) while surrounding healthy tissues are preserved from damage. If the motion itself can be accurately estimated, the treatment plan and/or delivery can be adapted to compensate.

Current methods for motion estimation rely either on invasive implanted fiducial markers, imperfect surrogate models based, for example, on external optical measurements or breathing traces, or expensive and rare systems like in-treatment MRI. These methods have limitations such as invasiveness, imperfect modelling, or high costs, underscoring the need for more efficient and accessible approaches to accurately estimate motion during radiation treatment. This research, in contrast, aims to achieve accurate motion prediction using only relatively low-quality, but almost universally available planar X-ray imaging. This is challenging since such images have poor soft tissue contrast and provide only 2D projections through the anatomy. However, our hypothesis suggests that, with strong priors in the form of learnt models for anatomical motion and image appearance, these images can provide sufficient information for accurate 3D motion reconstruction.

We initially proposed an end-to-end graph neural network (GNN) architecture aimed at learning mesh regression using a patient-specific template organ geometry and deep features extracted from kV images at arbitrary projection angles. However, this approach proved to be more time-consuming during training. As an alternative, a second framework was proposed, based on a self-attention convolutional neural network (CNN) architecture. This model focuses on learning mappings between deep semantic angle-dependent X-ray image features and the corresponding encoded deformation latent representations of deformed point clouds of the patient's organ geometry.

Both frameworks underwent quantitative testing on synthetic respiratory motion scenarios and qualitative assessment on in-treatment images obtained over a full scan series for liver cancer patients. For the first framework, the overall mean prediction errors on synthetic motion test datasets were 0.16±0.13 mm, 0.18±0.19 mm, 0.22±0.34 mm, and 0.12±0.11 mm, with mean peak prediction errors of 1.39 mm, 1.99 mm, 3.29 mm, and 1.16 mm. As for the second framework, the overall mean prediction errors on synthetic motion test datasets were 0.065±0.04 mm, 0.088±0.06 mm, 0.084±0.04 mm, and 0.059±0.04 mm, with mean peak prediction errors of 0.29 mm, 0.39 mm, 0.30 mm, and 0.25 mm.

List of Abbreviations

ABC	Active Breathing Control
AMM	Active Motion Mitigation
AP	Anterior-posterior
AE	Autoencoder
AI	Artificial Intelligence
ANN	Artificial Neural Network
CBCT	Cone Beam Computed Tomography
CNN	Convolutional Neural Network
CS	Compressed Sensing
CT	Computed Tomography
DIBH	Deep Inspiration Breath Hold
DIR	Deformable Image Registration
DRR	Digitally Reconstructed Radiograph
DVF	Deformation Vector Field
ECM	External Correlation Model
EHR	Electronic Health Record
EPID	Electronic Portal Imaging Device
FBP	Filtered-back-projection
FDK	Feldkamp, Davis and Kress
FM	Fiducial Marker
FPN	Feature Pooling Network
GAN	Generative Adversarial Network
GI	Gastrointestinal
GPU	Graphical Processing Units
GNN	Graph Neural Network
GTV	Gross Target Volume
IGRT	Image-guided Radiotherapy
IMRT	Intensity Modulated Radiation Therapy
ITV	Internal Target Volume
kV	kiloVoltage
linac	Linear Accelerators
ML	Machine Learning
MLC	Multi-leaf Collimator
mPD	Mean-projection-distance
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
MSE	Mean Square Error
mTRE	Mean-target-registration-error
MV	Megavoltage

NCC	Normalized cross-correlation
NTCP	Normal Tissue Complication Probability
OAR	Organ At Risk
PCA	Principal Component Analysis
PDF	Probability Density Function
PSNR	Peak Signal-to-noise Ratio
PTV	Planning Target Volume
ReLU	Rectified Linear Unit
RMSE	Root-mean-squared Error
ROI	Region of Interest
RPM	Real-time Position Management
RT	Radiotherapy
RTRT	Real-time Tracking Radiotherapy
SABR	Stereotactic Ablative Body Radiotherapy
SBRT	Stereotactic Body Radiotherapy
SGRT	Surface Guided Radiotherapy
SI	Superior-inferior
SRS	Stereotactic Radiosurgery
SSIM	Structural Similarity Index Measure
STD	Standard Deviation
TACE	Transarterial Chemoembolization
TCP	Tumour Control Probability
TI	Therapeutic Index
TRE	Target Registration Error
UQI	Universal Quality Index
US	Ultrasound System
VAE	Variational Autoencoder
VMAT	Volumetric Modulated Arc Therapy

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Radiotherapy (RT) has become a pillar of cancer treatment all over the world during the last few decades. Historically, RT involved large, square radiation fields delivered to a large anatomical region surrounding the tumour (target), leading to significant toxicity, and limited deliverable dose. Medical imaging has revolutionized RT, leading to precise, conformal radiation fields, optimized to a static single time-point representation of patient anatomy. This representation, however, does not account for patient motion, which may cause overdosing of organs-at-risk (OARs), or under-dosing of the tumour, leading to poorer outcomes for survival and post-treatment morbidity [1].

External beam radiation has become a standard of care in cancer RT clinics since it is used daily in numerous hospitals and health-care centres to determine the internal anatomical structure of organs and tumours during treatment planning and delivery [2]. Precision is critical for achieving tumour coverage while preserving surrounding

sensitive healthy tissues [1], as damage to normal tissues hinders the escalation of the dose to the desired therapeutic level in the gross target volume (GTV). In treatments with higher, but more precisely targeted doses (hypo-fractionated treatments) such as Stereotactic Ablative Body RT (SABR), unaccounted patient motion is yet more critical, and sometimes even prohibitive [3]. Hence, to fully exploit the potential of external beam radiation, tumour and organ movements must be addressed during the irradiation process, ensuring that more radiation is delivered to the target tumour while sparing OARs.

1.2 Clinical motivation: impact of motion on treatment efficacy

1.2.1 Periodic motion impact

Numerous strategies have been developed to manage breathing-induced (i.e. periodic) motions with external beam radiation techniques, such as SABR. These strategies can be broadly categorized into two approaches: passive and active motion mitigation (AMM) techniques [4, 1]. An example of a passive mitigation technique that has generally been used in radiation treatment to account for breathing motion is defining an internal-target-volume (ITV) (see Figure 1.1). The ITV includes a motion-encompassing safety margin, i.e., clinical target volume + internal margin to account for target motion [4]. The motion-encompassing margin is normally estimated based on 4D-CT data acquired at treatment planning and hence reflects an average estimate of the motion that corresponds with in-treatment motion with uncertain accuracy. These margins lead to inaccuracies during treatment and result in greater irradiation of nor-

mal tissues [5], meaning overall radiation intensity must be reduced, and treating the target tumour becomes more difficult.

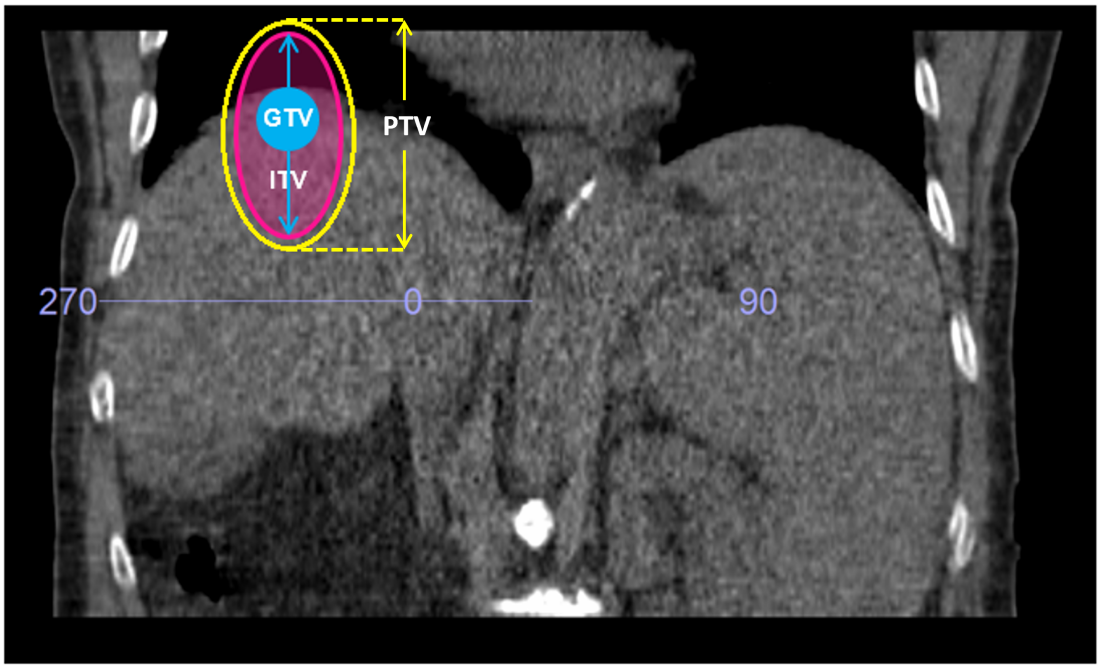


Figure 1.1: An illustration of target volumes accommodating respiratory motion. The visible or palpable extent of the tumour, known as the GTV, is depicted in blue, while the ITV, designed to ensure comprehensive coverage despite variations in position and shape, is represented in pink. The PTV is delineated in yellow and incorporates additional margins to address uncertainties in treatment delivery, such as setup errors and organ motion.

Clinicians also take into account a planning target volume (PTV) margin in addition to ITV, which is utilised to account for treatment-related patient positioning difficulties (i.e., ITV plus setup margin). This can be reduced by accurately delineating the ITV, which allows for dose escalation to the gross target volume (GTV). The aim is therefore to reduce or eliminate ITVs. To achieve this goal, the utilization of an AMM technique is required. This technique relies on real-time information about the tumour position, as it has the potential to significantly reduce the ITV margin and, consequently, mini-

mize the radiation dose to surrounding normal tissues.

Respiratory-gating and tracking are two popular AMM methods [5], although they have their own limitations. Respiratory-gating is relatively easy to implement, but it implies a longer time to deliver the specified dose because radiation is only delivered for a segment of the respiratory cycle [1]. In contrast, real-time tracking, which repositions and/or reshapes the radiation beam as the target moves, implies no prolongation of treatment sessions but is more technically challenging to realise. Moreover, its effectiveness can be limited by the time delay between detecting a change in target position and the system adjustment, resulting in a persistent lag in the system's response to the target position.

All active methods critically rely on real-time information on the tumour's position during treatment. Gating and tracking techniques often rely on implanted markers to track the target's mobility in real-time. These markers are invasive, and in any case, only provide information on specific locations (i.e., marker positions) inside tissues, rather than the target/OARs as a whole [6]. Moreover, implanting fiducial markers (FMs) may lead to organ inflammation due to infection, bleeding, displacement or migration during the treatment delivery [7, 8]. Therefore, techniques based on non-invasive imaging are preferred. Treatment systems integrating magnetic resonance imaging (MR-linac) arguably provide an excellent basis for this [9] in the form of real-time in-treatment images that are radiation-free and have good soft tissue contrast and resolution [10]. However, current MR-linacs provide only orthogonal pairs of 2D slices rather than true 3D images and hence do not directly enable visualisation of the whole 3D geometry of a target tumour region and surrounding OARs. More importantly, such systems are expensive and rare, meaning very few patients currently can access them.

In contrast, most conventional linacs are equipped with on-board kV (kilovoltage) X-ray imaging, and such systems will inevitably be used to treat most patients; techniques that can recover anatomical motion from such images are therefore attractive.

1.2.2 Therapeutic Index with Motion impact

The purpose of this research is to enhance the probability of curing cancer by maximising tumour control probability (TCP) while minimising normal tissue complication probability (NTCP). The relationship between TCP and NTCP is shown in Figure 1.2.

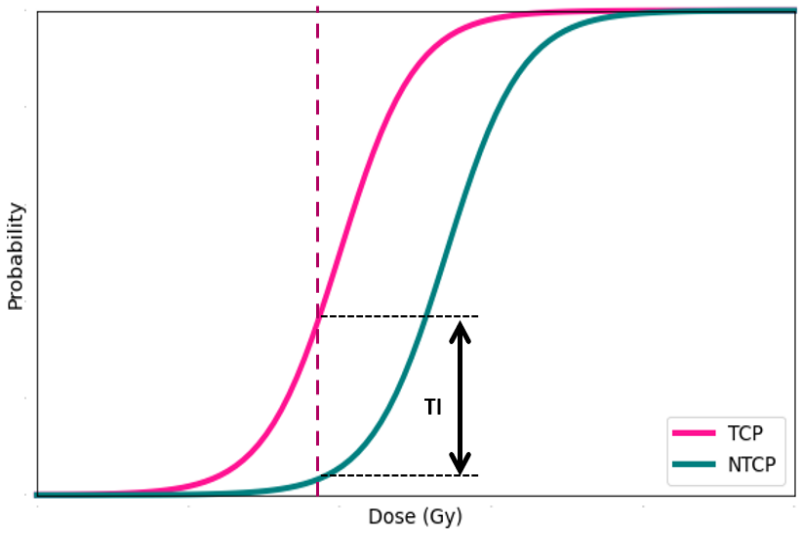


Figure 1.2: The dose-response curve of TCP represented in pink colour and NTCP represented in green colour with respect to radiation dose in conventional RT. Sparing normal tissues shifts the NTCP curve to the right, allowing a lower incidence of normal tissue damage for the same dose.

It is not feasible to select a radiation dose that would completely cure cancer while producing no side effects to normal tissues since the separation of these two curves is governed mostly by biological factors such as tumour characteristics, patient physiol-

ogy, and genetic variability [11] and therefore clinicians have no control over it. The ideal circumstance is to conduct a clinically gentle intervention. To achieve this, the therapeutic index (TI), which is the difference between the likelihood of curing cancer and the probability of causing an unacceptable side effect, can be computed. The graph shown in Figure 1.2 depicts the dose-response curve for a conventional RT distribution, indicating that a relatively low TCP value can be achieved while also having a low NTCP value.

While the ideal TI in RT is set at one, clinicians often encounter challenges in reaching this optimum value due to the inherent limitations imposed by biological factors as described above. To address this challenge, the solution involves decoupling doses, enabling clinicians to administer a high radiation dosage to the tumour while minimizing exposure to normal tissues. Conformal RT techniques, such as SABR, prove valuable in this context by enhancing TCP while keeping NTCP low, thereby increasing the TI. The graph in Figure 1.3 illustrates a favourable scenario where patients experience significantly greater TCP for the same level of toxicity.

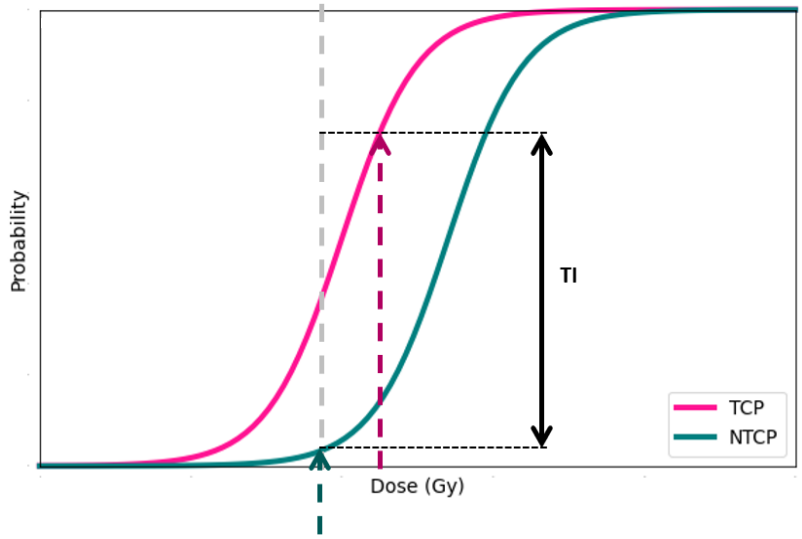


Figure 1.3: Graphical visualization illustrates the increase of the TI by using a conformal radiation technique. This enables a comparable level of NTCP, akin to the conventional approach, but at a higher dose.

However, when there is motion, the healthy tissue located in the low-dose zone may migrate into the treatment field, leading to the delivery of a higher radiation dose to the healthy tissue. This, in turn, leads to patients enduring intolerable levels of toxicity. Another scenario arises when the tumour moves away from the radiation field, resulting in a reduced radiation dose to the tumour. Despite the consistent toxicity status, with the healthy tissue remaining stationary while the tumour moves, the tumour receives a suboptimal radiation dose, resulting in a lower TCP value. This necessitates a reduction in the prescribed dose to bring the NTCP back to an acceptable level. Hence, patients may not fully benefit from conformal radiation treatments like SABR due to challenges associated with motion. This, in turn, poses challenges for clinicians in delivering an optimal radiation dosage to patients, ultimately influencing the overall outcomes for patients.

This project aims to ameliorate patient motion effects and improve TI via AMM, en-

abling clinicians to reduce the NTCP to acceptable levels, thus allowing for larger prescription doses that will result in better outcomes for the patient. This will then enable clinicians to adapt therapies for diverse patient cohorts.

1.2.3 Target Under-Coverage

Suppose that the ITV always encompasses the complete motion of the GTV, despite the fact that it is a snapshot taken during the simulated CT scan and that the GTV may spend some time outside the ITV. If the tumour is always located inside the planned radiation dosage (i.e., within the pink region in Figure 1.4), a TCP of 85-90% can be achieved. However, if the tumour escapes the radiation field even 20% of the time, the TCP falls nearly 60%. This implies that tumour moves outside the radiation field result in a loss of almost 30% of TCP [12]. As shown in Figure 1.4, decreasing the effective radiation dosage has a significant impact on TCP.

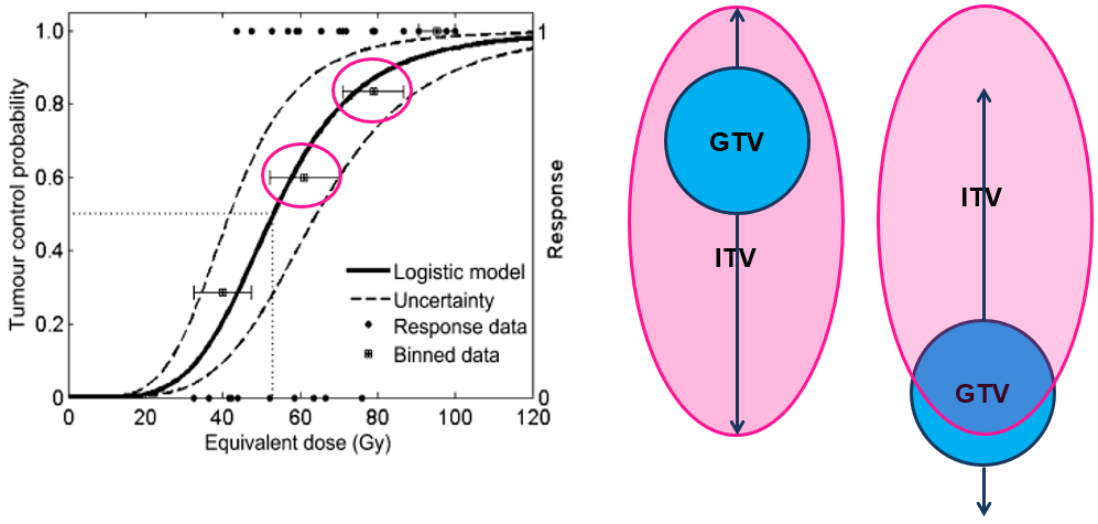


Figure 1.4: Target Coverage during Treatment

Ideally, radiation delivery should be designed to follow a moving target if the tu-

mour/organ movement can be tracked continuously in real-time without markers. This would eliminate the requirement for a tumour-motion margin, resulting in less irradiation of normal tissue.

1.3 Aims

The overall aim of the work was to develop artificial intelligence (AI)-based patient-specific motion modelling techniques that enable the prediction of internal anatomical motion from highly sparse, low-quality in-treatment kV projection images. To this end, three main research streams were pursued:

- Development of techniques for generating synthetic training data, as model training necessitates paired sets of organ motion instances and corresponding kV images. This involved synthesizing data based on patient-specific motion patterns derived from 4D-CT images.
- Development of an end-to-end GNN-based approach to predict 3D volumetric organ shape deformation from a single in-treatment kV planar X-ray image acquired at any arbitrary projection angle.
- Development of a simpler CNN approach that exploits learnt patterns of tissue motions by extracting semantic information from single in-treatment kV planar X-ray images.

1.4 Potential Challenges

It is difficult to predict complex internal anatomical motion from highly sparse kV planar X-ray image details. One primary difficulty arising from these images involves instantaneous projection over the patient volume, compressing volumetric information along projection directions, i.e., in the direction of the beam axis. Additionally, kV image contrast is very poor in soft tissue regions, which in any given projection can be obscured by bony anatomical features. Although deep learning techniques have been shown to enable up-sampling and image synthesis based on low-contrast or low-resolution medical images, combining this requirement with inferring a 3D geometry from a 2D image poses a significant challenge. To overcome this challenge, we use the corresponding patient-specific organ as a volumetric 3D mesh prior which is extracted from the reference CT volume.

1.5 Contributions

Two distinct patient-specific deep learning techniques have been developed for reconstructing 3D volumetric organ models from an arbitrary angled single-view kV X-ray.

In the first phase of the research, we proposed a model that learns mesh regression from a patient-specific template and deep features extracted from kV images at arbitrary projection angles. A 2D-CNN encoder extracts image features, and four feature pooling networks fuse these features to the 3D template organ mesh. A ResNet-based graph attention network then deforms the feature-encoded mesh. The first phase of the research offered the following contributions:

- An end-to-end deep learning technique that integrates a convolutional neural

network (CNN) image encoder and a graph-attention network through learnable feature pooling networks (FPNs) for reconstructing 3D volumetric organ models from an arbitrary gantry-angled single-view kV planar x-ray image.

- The arbitrary projection angle information is incorporated via an additional channel to the input image to extract angle-dependent features so that the model can reconstruct the 3D anatomy from kV images acquired at any projection angle.
- The image features are fused into the 3D mesh space through four learnable FPNs where each FPN is associated with its corresponding convolutional layer in the encoder to extract hierarchical features for each vertex. This enables non-trainable components (i.e. the vertex projection operation) to be eliminated from the model architecture, making it end-to-end trainable.
- To the best of our knowledge, this is the first time a deep learning framework has been used to reconstruct volumetric 3D organ models from an arbitrary gantry-angled single-view medical image.

In the second phase of the research, an attention-based CNN model was proposed that enables estimating lower-dimensional representations of the vertex-wise displacement field for the motion of internal anatomy. This second approach was introduced to address the time-consuming nature of the first approach based on the GNN framework. In contrast, the CNN-based method efficiently learns a mapping from kV image-derived features to the lower-dimensional representation of the vertex-wise displacement field of deformed meshes. This reduction in learnable model parameters results in a more efficient and robust approach, with the CNN-based method taking only around 4 msec per input image during inference, compared to the GNN method's requirement of approximately 27 msec.

Moreover, a CycleGAN, conditioned on the projection angle, was developed and trained using an unpaired set of real kV X-rays and DRRs as part of the synthetic data generation process. The purpose of this model is to facilitate the style transfer from real kV images to DRRs. The motivation behind adopting this approach is to address the inherent limitations of DRRs, which lack the scatter properties and noise characteristics found in real kV X-ray images. This model was trained for each case individually since the field of view (FOV) acquisition varies from patient to patient.

1.6 Research Articles and other outputs

In addition to the contributions listed above, the following research papers have been submitted for publication so far.

- Deep-Motion-Net: GNN-based volumetric organ shape reconstruction from single-view 2D projections, submitted to Medical Image Analysis Journal.

Moreover, the following research papers have been prepared for publication.

- An attention-based CNN framework for volumetric organ shape reconstruction from single-view 2D projections.

In addition to that, I showcased my research in:

- AI workshop/AI theme category at the Annual Meeting for Radiation Research (ARR) in 2021.
- A poster abstract presented in CRUK-ARR Radiation Research Conference in Glasgow 2023.
- A poster abstract presented in CRUK RadNet PhD & Postdoc Symposium in

London 2023.

1.7 Organization of the Thesis

The content of this thesis is organised as follows.

- **Chapter 2** provides a systematic review of motion monitoring and management in RT. This chapter encompasses an exploration of X-ray-based motion monitoring methods, surface-guided methods, and hybrid, hypo-fractionated and MRI treatment techniques for motion mitigation. Additionally, it delves into recent advancements that leverage machine learning and deep learning models to estimate internal anatomical motions.

- **Chapter 3** presents a patient-specific end-to-end deep learning model to build the complex relationship between deformations of the target anatomy and the corresponding appearance of that anatomy in X-ray images acquired at any arbitrary projection angle. The evaluation outcomes are subjected to both quantitative and qualitative analysis, utilizing a synthetic motion dataset, and in-treatment images obtained throughout the full scan series for four liver cancer patients respectively. This chapter is derived from my paper "Deep-Motion-Net: GNN-based volumetric organ shape reconstruction from single-view 2D projections" submitted to the Medical Image Analysis Journal.

- **Chapter 4** presents a novel self-attention-based CNN framework for estimating the relationship between the lower-dimensional representation of 3D organ deformations and single in-treatment kV planar X-ray images. This proposed method addresses the high time cost of the model training process highlighted

in the previous model discussed in Chapter 3. This chapter is derived from my paper ”An attention-based CNN framework for volumetric organ shape reconstruction from single-view 2D projections”.

- **Chapter 5** provides a concise overview of the research outcomes, offering insights into the practical application of my findings. It also includes the limitations identified during the study and outlines potential future work.

Chapter 2

Literature Review

This chapter provides a comprehensive exploration of existing approaches related to motion monitoring and management in RT, systematically divided into six main sections. Section 2.1 describes kV and megavoltage (MV) X-ray approaches used during the past few decades including their limitations. Section 2.2 describes Surface-guided approaches. Section 2.3 explains hybrid methods used during the past few decades including their limitations. Section 2.4 and 2.5 explicate the extended systems that can be used for motion monitoring with electromagnetic markers and ultrasound-based approaches, respectively. Section 2.6 describes the magnetic resonance imaging for motion management and its limitations. (For a detailed review of topics outlined in sections 2.1-2.6, see Bertholet et al. [13]). Finally, section 2.7 describes learning-based techniques, and comprises four parts, namely machine learning techniques (particularly deep neural networks that were used for respiratory motion management tasks), deep learning-based image registration techniques, surrogate-based motion models and shape reconstruction from single-view projections. Finally, Section 2.8 offers a sum-

mary of this chapter.

2.1 X-ray-based motion monitoring

Image-guided RT (IGRT) based on kV and MV X-ray imaging is progressively being used for target tumour localization and patient setup in RT treatments. These techniques play an essential role in delivering a highly conformed dose to the target with precision. MV X-ray imaging gives poor soft-tissue contrast compared to the kV X-ray imaging techniques. The main issue, apart from poor soft tissue contrast, is that the MV FOV is also very truncated and irregular since the collimator leaves shape the beam to conform to the target. Electronic Portal Imaging Devices (EPIDs) and kV X-ray devices that are integrated into the treatment unit are the most commonly used X-ray imaging devices. These strategies come in various hardware configurations of monoscopic or stereoscopic imaging and it is possible to combine these techniques with external monitoring as well. To extract the target position from the planar image or sequence of images, image processing algorithms are used by these image-guided approaches. Moreover, the acquisition and the processing time of the image directly impact the latency of these methods [14].

2.1.1 Implanted fiducial markers

Due to the poor soft-tissue contrast, one key problem in X-ray imaging is the difficulty in detecting the target tumour position compared to the nearby sensitive tissues. To overcome this issue, radiologists surgically implant the high-contrast fiducial markers (FMs) close to the target tumour for the image guidance of pretreatment. This technique helps pinpoint the location of tumours with much higher accuracy and the

ability to deliver the homogeneous dosage to the tumour while keeping the radiation to the critical surrounding tissues or organs to a minimum. These implantations can be done through needle-punctures into the organs, for instance, lungs, pancreas, liver and prostate [15, 16, 17].

It is also possible to implant surrogates using the endoscope into or close to the gastrointestinal (GI) tract [18, 19] whereas the surgical implantation approach is used to monitor the paraspinal and spinal lesions [15]. Endovascular platinum coils were implanted in [20] to mark intrapulmonary lesions by placing them in branches of the pulmonary artery that are very close proximity to the target tumour. Hepatocellular carcinoma was detected using Coiled FMs as an internal surrogate in [21] while a thin FM namely Gold Anchor implanted for 621 patients who have prostate cancer to reduce the infections due to conventional FMs associated with 17G-18G needles [22]. Later, an experiment using the airway-implanted FMs and an external surrogate for 28 patients with lung tumours was carried out by Willmann et al. [23]. Comparing tumour motion in the anterior-posterior (AP) and superior-inferior (SI) directions, the authors revealed that internal FMs appear to be more accurate predictors of lung tumour motion than the exterior surrogate. Akasaka et al. [24] conducted a study with 230 lung cancer patients to investigate the relationship between ITV margin and FM position using the SABR technique. Very recently, Joon et al. [25] conducted a clinical trial to compare the effectiveness of gold and polymer FMs in the treatment of prostate cancer with 28 patients.

Traditional FMs may result in complications due to migration or displacement and therefore Rose et al. [26] used a liquid FM called Lipiodol to detect margins of the target tumour and to discover small peripheral malignancies for lung lesions. Moreover,

there is evidence indicating that residual lipiodol on imaging can serve as a surrogate marker for individuals who have undergone transarterial chemoembolization (TACE). This approach helps avoid potential issues associated with placing FMs. [27].

The automatic segmentation of FMs must be done in real-time for any treatment based on intra-fraction monitoring. By tracking the movement of the FMs in the images, it is possible to adjust the radiation delivery in real-time. However, this is difficult for MV X-rays due to their lower contrast compared to kV X-rays [28]. This further indicates that the target and FMs may not be easily distinguishable from surrounding tissue in MV X-rays, making it difficult to accurately segment the target tissue. To enhance the contrast level of the MV X-rays, Short-arc digital tomosynthesis (SA-DTS) was used in [29] and then joined with kV X-rays that were acquired in orthogonal directions for monitoring the 3D motion of the target tumour. A simple parametric template was used in [30, 31, 32, 33] to segment the spherical or cylindrical FMs in real-time from MV or kV X-ray projections. The template is typically designed to match the shape and size of the FM and is then applied to the X-ray image to extract the position and orientation of the FM in real-time.

Complex templates can be used for accurately segmenting the FMs that exhibit arbitrary shapes. One way to generate such templates is by acquiring breath-hold CT [34] or cone-beam CT (CBCT) projections before treatment [35]. These templates can then be used to segment the FMs in real-time during treatment, allowing for more accurate tracking of their position and orientation. However, the generation of these complex templates may require additional time and resources, and may not always be necessary depending on the shape and size of the FM. Template mapping needs to match object shape that dramatically varies for various implantations and angles of projec-

tion. Therefore, this requires a considerable number of templates to cover numerous circumstances. However, to reduce the computational load, template-based methods are usually forced to use a minor number of templates since they use an exhaustive search in the region of interest (ROI). To solve this issue, a template-free approach was introduced by Lin et al. [28] and the authors used discriminant analysis to segment implanted FMs, and for the sequential tracking, they used mean-shift feature space analysis. Another template-free approach was proposed by Wan et al. [36] to segment the FMs based on dynamic programming for the projection images of CBCT. Then, it was used to adjust the position of the couch optimally for the treatment and/or bounds of the gating window.

The toxicity risk, complications and additional costs are often associated with FM implantation. The most common complication is pneumothorax due to percutaneous implantation in the lung as mentioned in [37] where the authors identified 20 pneumothorax cases among 44 lung implantations. Another example is urinary tract infection due to the implantation of trans-rectal FMs in the prostate [22]. These health complications due to FM implantation can be reduced by utilizing thin FMs that require a small needle stick [22]. The insertion of FMs into the body takes considerable time, thereby causing delays in the treatment delivery. Moreover, this FM insertion may lead to organ inflammation and FM displacement or migration during the treatment. Another complication is that the tumour position changes with respect to the implanted markers due to the deformations of the target tissue. Some clinical approaches for tumour localization depend on internal anatomical surrogates such as chest wall and diaphragms, however, the main limitation of this approach is the accuracy given the poor correlation of the degree of motion between the target tumours and the anatomical surrogates [38].

2.1.2 Stereoscopic X-ray imaging methods

Stereoscopic X-ray imaging is capable of providing 3D spatial information from 2D projection X-ray images. However, this X-ray imaging technique requires additional tools such as the CyberKnife system, real-time tracking RT system (RTRT), Vero system, etc. These systems are designed to provide high-quality, real-time imaging that can be used to track the position and movement of the target tissue during treatment. Stereoscopic imaging, facilitated by multiple X-ray images taken from different angles, allows for the creation of a 3D reconstruction of the target area. This, in turn, enables more precise and accurate targeting of the radiation.

The CyberKnife device was invented in the early '90s by applying stereoscopic principles. It serves as an instrument for markerless tumour localization in frameless stereotactic radiosurgery (SRS) and stereotactic body RT (SBRT) [39]. This is the first clinical system with a linear accelerator (linac) for real-time motion tracking and prediction. In early 2000, this system was improved by using FM implantation for treating extra-cranial tumours in the pancreas and spinal cord [40]. The cyberknife system comprises two flat-panel detectors that were placed as opposed to each other and mounted on the floor to capture the images of the tumour from different angles, kV X-ray imaging sources were mounted on the ceiling, allowing for accurate targeting of tumours. The system also includes an MV-linac, mounted on a robotic arm, which delivers high-energy radiation beams to the tumour. This is the first machine that was able to follow the target tumour motion and tracking since the treatment beam can rotate six degrees of freedom by re-aligning the robotic linac. However, the main limitation of this system is that it is inadequate to resolve the motion of the respiration since X-rays can only be acquired every ten or twenty-second period of time during the

treatment delivery. This system has been used to monitor the motion of prostrates in the recent past during SBRT treatment [41, 42, 43, 44, 45] even inadequate to resolve breathing motion. However, this system can only handle tiny tumour volumes and suffers from system latency due to repeated verifications before each radiation delivery through the beam.

The RTRT system [46] was designed to ensure that the patient's tumour is within the treatment field at all times during the radiation delivery. Employing high-frequency stereoscopic X-ray imaging, this system provides continuous monitoring of the tumour's position and motion. It possesses the capability to dynamically adjust the radiation beam's position to compensate for any tumour movement. The initial system has recognized the location of a 2 mm gold FM within the patient's body with a 1 mm accuracy for every 0.03 seconds during radiation delivery by utilizing synchronized linac. This system consists of four kV X-ray imaging sources that were placed in the corners of the floor with corresponding detectors that were mounted on the ceiling. The radiation was stopped by the linac if the implanted gold FM was not within the range of the gating window and radiation was delivered if the FM was within the gating window range compared to the planned position. The pulses from the kV X-ray imaging and linac systems were synchronized using MV scatter-free kV X-ray images. Hanazawa et al. [21] developed a simple template-based matching algorithm to acquire pairs of 30 kV X-ray images per second for detecting a Visicoil or spherical FM position. The RTRT system is associated with a high monitoring rate and therefore it has extensively permitted for detection of the motion of the target tumour within numerous anatomical sites [15, 47]. Shiinoki et al. [48] proposed an approach for the respiratory-gated RT verification with the SyncTrax system which was similar to the RTRT system by

using cine EPID images and a log file. The authors used internal surrogates (inter and intra-fractional variations of the implanted FM) to evaluate the gating accuracy.

Kamino et al. [49] described a system called Vero that consists of a gimbals-supported small linac head with an o-ring gantry to deliver treatments by precisely locating moving tumour targets in real-time. It utilizes a robotic couch and a gimbaled treatment head, which allows for six degrees of freedom in patient positioning and beam delivery. This system is equipped with two kV X-ray sources that are orthogonal to each other and paired with opposite flat panel detectors. These components are attached to the o-ring gantry at a 45-degree angle relative to the radiation beam. To follow the respiratory motion of the target tumour, the treatment beam of this system used the skew angle of the gantry along with the tilt and pan movement of the two gimbals. This system also relies on the external correlation model (ECM) capability between a superficial surrogate motion and internal anatomical target motion to predict the target tumour position. This system is no longer accessible worldwide.

Mori et al. [50] designed a markerless stereoscopic monitoring approach to treat both liver and lung cancer patients by detecting the moving target tumour. This system acquired an image sequence for a patient during the respiration cycle. The authors used a machine learning-based multi-template matching algorithm where learning is conducted by using the pretreatment images for each patient. The authors evaluated patient setup accuracy, radiation dosage, gating positional accuracy and workflow of the treatment delivery. Patient setup accuracy was computed by using 2D to 3D image registration technique between reference DRR and flat-panel detector images.

2.1.3 Other X-ray-based approaches

Using MV projections during dynamic tumour tracking, Serpa et al. [51] suggested a dense-feature-based technique for estimating the mobility of the soft tissues. The authors evaluated the performance of their algorithm by applying it to fluoroscopic sequences acquired at ~2 Hz for a dynamic phantom and two lung cancer patients treated with the SABR system. For the dynamic phantom, the root-mean-square error (RMSE) was less than 1.2 mm, whereas for the clinical dataset, it was less than 1.8 mm. Roa et al. [52] recently conducted a research to investigate the dosimetric impact on the lungs with a kV X-ray beam from an infrared fluoroscope to deliver low-dose RT using Monte Carlo simulations and an acrylic phantom.

Fergusen et al. [53] proposed a markerless tumour monitoring algorithm which is based on MV cine EPID images for lungs derived from a dynamic thorax phantom. During the delivery of radiation, dynamic phantom images were acquired for several lung SABR breathing traces and a sample patient data set. The phantom data had a tracking error of 1.34 mm, while the patient data had a tracking error of 0.68 mm. Later, Bruin et al. [54] proposed a marker-less approach to track the lung tumours in real-time using kV-based SABR during VMAT. A series of planar kV images was acquired at 7 Hz during treatment delivery from a 3D phantom which comprises three lung tumour targets. The authors used the inspiration phase in 4D-CT to generate 2D reference templates for different gantry angles. Normalized cross-correlation was used to match the kV X-rays and templates to recognize the tumour positions. The third dimension was recognized by the triangularization of 2D-matched projections. In 92% and 96% of the kV projections for the phantom's targets, 1 and 2, 3D findings within the 2 mm range of the known position were present, respectively. This percentage

plummeted to 80% for target 3. Recently, Mueller et al. [55] proposed a markerless tumour tracking technique utilizing intra-fractional kVs to perform a trial for 30 lung cancer patients. The treatment is interrupted by the clinician if the mean lung tumour position shifts by more than 3 mm, and it is then resumed after adjusting the treatment couch to account for the shift. This technique is considered effective if the tracking accuracy is less than 3 mm in each dimension for more than 80% of the treatment time.

2.2 Surface guided motion monitoring

Surface imaging is a technique which has the ability to track the patient's skin surface through advanced 3D camera technologies. This employs real-time optical imaging techniques such as structured-light-imaging [56], time-of-flight [57], laser scanning [58] and stereo-vision [59] to generate a 3D surface of the patient. The primary benefit of surface-guided radiation lies in the fact that it does not involve ionizing radiation. This approach is progressively used for motion management in radiation treatments. During the surface-guided RT (SGRT) treatment, a camera is used to position and monitor the external surface of the patient to ensure whether the radiation is accurately targeted.

2.2.1 Infrared marker-based approaches

Stereoscopic in-room cameras are acting as an external surrogate for the target position to detect the position of the IR reflectors [60, 61]. Besides, systems such as real-time position management (RPM) and ExacTrac 6D [62] can be utilized for extracranial respiratory-gated RT. The geometric accuracy of RPM was computed by using FM

trajectories for pancreas, lung and liver patients [63] and deep-inspiration breath-hold (DIBH) used to treat lung patients with visual feedback [64]. In 2018, Fassi et al. [65] reported a 5.8 mm error for 3D median residual set-up compared to the kV images with implanted clips utilizing multiple reflectors for the treatments of RPM-guided left-sided breast DIBH. The treatment can be interrupted and the position of the patient can be changed based on the acquired volumetric imaging when the position of the internal anatomical structure gets changed [66].

The advantages of using infrared markers for motion monitoring in radiotherapy include providing a non-invasive and non-ionizing method to monitor patient motion during treatment sessions. This approach reduces discomfort and eliminates the risk associated with additional ionizing radiation exposure for patients. However, there are several limitations associated with using infrared marker systems for motion monitoring in radiotherapy. These systems typically require an unobstructed line of sight between the markers and monitoring cameras, which can restrict their use in certain treatment setups where obstructions or patient positioning may interfere with monitoring accuracy. Achieving and maintaining accurate calibration and positioning of infrared markers can be challenging, often requiring regular adjustments to maintain accuracy. Additionally, external factors such as ambient lighting and reflections can influence the performance of infrared marker systems, potentially affecting the accuracy of motion monitoring. Furthermore, patients must remain still and cooperative during treatment to ensure accurate data capture with infrared markers, which can be challenging for some individuals. Lastly, it's important to note that infrared markers primarily capture surface motion and may not provide comprehensive information about internal organ motion due to imperfect correlation between surface and internal

motion dynamics [67].

2.2.2 Optical surface-based approaches

To map the surface of a given patient, one or more high-definition cameras are used in the optical surface monitoring systems. Catalyst and AlignRT are such kinds of systems that use two and three room-mounted cameras, respectively, for estimating the six degrees of freedom organ motion by projecting the structured light patterns [61]. The reference surface which is obtained through a simulated CT can be used to compare the patient surface that is detected in real-time during the treatment delivery. Image registration techniques have been used to register the subsets of the surface with respect to the reference surface to report the real-time rotation and translation of the patient [68, 69, 70]. The beam-hold is triggered automatically by certain integrated systems, for instance, AlignRT when there exists a mismatch between the reference surface and the current surface. The patient adjustment is also possible with this system for optimal matching by using immediate in-room feedback.

Surface guidance for monitoring the intra-fraction motion was primarily used for DIBH breast treatments [71, 72, 73]. Moreover, AlignRT employs active stereovision technology to track patient movement with precision to the sub-millimetre level [74, 75, 76, 77]. Recently, Sorgato et al. [78] conducted a study to evaluate the precision of the AlignRT technique in identifying and measuring oedema during RT for breast cancer using water-equivalent boluses and a female torso phantom. Recently, a clinical workflow based on an SGRT procedure was proposed by Li et al. [79] for treating breast cancer patients using DIBH. During simulation, both free-breathing and DIBH CT scans were obtained to measure the anterior surface displacement and then the au-

thors performed an alignment from free-breathing to DIBH to obtain the residual setup errors. This research was conducted using 26 optical surface imaging systems in nine clinical centres.

Optical surface monitoring techniques are non-invasive and rely on cameras to capture surface movements without requiring implanted markers. These methods avoid the use of ionizing radiation by utilizing cameras and sensors to detect surface points or features on the patient's skin. However, the accuracy of surface monitoring depends on certain factors such as the colour of the patient's clothes, skin tone, the visible light and the reflectivity from in-room lighting [61].

2.2.3 Other surrogate-driven approaches

The use of a spirometer allows for the measurement of air volume within the lungs at a particular moment in time. To ensure accurate readings, a nose clip is worn by the patient while undergoing the breathing procedure [80]. Moreover, the incorporation of a scissor valve is advantageous for regulating the air volume at a desired level, as it aids in enforcing a breath-hold to reduce the motion of the target area. This concept is called active breathing control (ABC) [81] and is particularly relevant in treatments where respiratory motion can lead to inaccuracies in delivering radiation. This has been used for the lung [82], breast cancers [83, 84, 85], and liver [86, 87] cancer patients. Despite its advantages, there are several drawbacks including the necessity for coaching sessions, ensuring patient compliance, and fostering effective communication between the patient and the radiologist since some individuals may find it challenging to consistently hold their breath, leading to variations in treatment sessions. Moreover, using this technique can lengthen the overall treatment time and therefore, patients

must practise breath-holding techniques, and each treatment session may last longer due to the need for precise coordination [88]. This approach is particularly effective for analyzing cancers in thoracic and abdominal regions where respiratory motion is significant. In cases where breath-holding is not feasible or where there is minimal respiratory impact on tumour position, this method may not provide significant benefits.

2.3 Hybrid motion monitoring approaches

The use of respiratory monitoring may prove inadequate for accurately detecting the position of internal target tumours, as it has been identified as a suboptimal surrogate [89]. This implies that relying solely on respiration-based tracking methods may not provide the precision required to effectively monitor and locate internal target tumours during motion monitoring. To overcome this issue, one solution that scientists investigated was to develop hybrid monitoring techniques specifically by combining the sparse imaging approaches with respiratory monitoring for internal tumour motion estimation over time.

2.3.1 Synchrony systems

Ozhasoglu et al. [90] proposed an approach called Synchrony by modifying the existing CyberKnife system [39] to track the real-time organ motion in three-dimensional space due to respiration. This system comprises light-emitting diode (LED) markers and three sets of cameras that were mounted in the ceiling in addition to the x-ray kV imaging and the robotic linac system of the CyberKnife. Moreover, this system helps manage the robotic arm to move the radiation beam progressively to such an extent that the beam consistently stays lined up with the tumour object by using external FMs. In

2009, Hoogeman et al. [91] used the Synchrony system to analyze lung cancer patients by taking the correlation error between the external breathing motion and the internal tumour motion utilizing the intra-treatment images. The SI direction demonstrated mean errors ranging from 0.2 to 1.9 mm, whereas left-right (LR) directions and AP possessed mean errors ranging from 0.1 to 1.9 mm and 0.2 to 2.5 mm, respectively. Later, Bibault et al. [92] conducted research based on fiducial-free lung tumour estimation for 51 patients using the Synchrony system. The authors achieved the overall survival rate was 85.5% and 79.4% at one year and two years respectively whereas the actuarial local control rate was 92% and 86% at one year and two years respectively.

Ferris et al. [93] conducted a study focused on monitoring and synchronizing 3D respiratory motion with radiation delivery, investigating various phantom motions using the motion Synchrony system on the Radixact for helical tomotherapy. To capture motion, LEDs were strategically placed on the patient's chest. In this research, 4D-CT scans were obtained from 13 subjects to formulate helical plans. This study achieved an RMSE of less than 1.5 mm between the programmed phantom positions and the Synchrony-modeled positions. In a related study, Tse et al. [94] recently conducted an assessment of this system's accuracy, employing a patient-specific breathing pattern with respiratory phase shifts. The authors observed that as the degree of phase shifts increased, tracking errors also escalated.

2.3.2 ExacTrac systems

Willoughby et al. [95] introduced an approach by using implanted FMs to deliver gated treatment for the target tumour or OAR localization. This system, ExacTrac, typically employs a combination of imaging modalities, such as infrared cameras or X-rays,

to continuously monitor and verify the position of the target region and OARs. This was designed to address issues such as accurate alignment during patient setup and the need for on-the-fly adjustments to ensure optimal targeting during treatment. The breathing pattern of the patient was extracted from infrared reflectors and acted as the gating signals. To facilitate automated couch adjustments, a strategically positioned infrared reflective star on the couch is employed. Additionally, an array of five to seven external infrared reflective markers is placed atop the patient, detected through infrared cameras mounted in the ceiling. During treatment, a pair of X-ray images are acquired when the reference gating level aligns with the external signal. Subsequently, a comparison is made between the 3D triangulated position of FMs and their respective reference positions. Moreover, if a discrepancy surpassing a predetermined tolerance is identified, the radiation beam is deactivated, and the couch position is dynamically adjusted to ensure alignment with the treatment plan. Later, Jin et al. [96] extended this approach by utilizing the kV X-ray imaging system to allow six degrees of freedom tumour localization. Recently, this technique extended by combining stereoscopic X-ray, optical surface, and thermal tracking in a single system (ExacTrac Dynamic) [62, 97] to reduce the limitations such as misalignment of live surface and reference surface.

2.3.3 Vero system-based approaches

The Vero system [49] is equipped with real-time imaging tools, including CBCT and fluoroscopy, allowing clinicians to visualize the tumour and surrounding anatomy immediately before and during treatment. In this system, an implanted FM becomes the focal point as the radiation beam gracefully orbits around the centre of gravity of the linac assembly. This system visually presented a tolerance radius of 3 mm as the ROI

for the position of estimated FM and if the tolerance exceeds a certain threshold then the radiologists have a chance to terminate the session. Depuydt et al. [98] conducted research for a group of ten lungs and liver SBRT patients utilizing the Vero SBRT gimbaled linac system for the first time to monitor the moving tumours in real-time. In 2013, Akimoto et al. [99] conducted a very similar analysis by considering 110 log files for 10 lung cancer patients and the authors recommended updating the model often to avoid drift-related errors. Later, Orecchia et al. [100] evaluated the radiation toxicity and feasibility of this system by utilizing a cohort of 789 cancer patients with 957 lesions, observing an acceptable level of acute toxicity.

2.3.4 Other hybrid approaches

Berbeco et al. [101] proposed an approach to predict the motion of the lung tumours by utilizing the optical Anzai belt and RTRT. This Anzai sensor belt was placed around the patient's abdomen to track the breathing signal and the placement of infrared reflectors on the treatment couch was tracked with an infrared camera. The authors evaluated the residual motions that were treated with respiratory gating by using eight lung cancer patients. However, this snug fit of the Anzai belt and the need to follow specific breathing instructions can indeed cause discomfort for some patients. Patients may find it challenging to maintain the required breathing pattern consistently throughout the treatment session, especially if they experience discomfort or difficulty with the breathing instructions[102].

Bertholet et al. [103] developed a hybrid approach, namely COSMIK, to monitor the real-time intra-fraction motion of the target tumours. This approach combines the linac system with sparse and optical monoscopic imaging techniques along with kV X-rays.

This system involves an auto-segmentation method for implanted FMs in pre-treatment CBCT projections [35]. To estimate the external 3D trajectories of FMs, the authors utilized the Gaussian distribution [104]. These trajectories are used for setting up the patients automatically and to fit an augmented linear ECM [105]. The authors used a continuous external signal from the ECM to estimate the positions of the internal FMs during the treatment time. Moreover, the authors used phantom-based simulations to validate this system. Recently, Ravkilde et al. [106] and Skouboe et al. [107] combined this COSMIK approach together with the reconstruction of the 4D tumour dose in real-time for online treatment verification during RT delivery.

Amoush et al. [108] conducted a study to analyze the impact of intra-fraction motion on breast cancer by using a two-hybrid approach namely a two-isocenter conventional technique versus a Single-isocenter hybrid IMRT technique. Later Liang at al. [109] proposed a robust optimization approach in IMRT based on a skin flashing to detect the position of the target tumour due to respiratory movements with five breast cancer patients.

Xiong et al. [110] conducted a study using MRI-linac with gating to monitor the intra-fractional motion of the prostate and its dosimetric impact. The authors used 174 sagittal 2D cine-MRI fractions from 10 patients for this study. With reference to the centroid position of the gating boundary, the mean prostate motion without gating was 0.6±1.0 mm and 0.0±0.6 mm in AP and SI direction, respectively.

Recently, pencil beam proton treatment with respiratory gating was used by Nanakali et al. [111] to track the movements of the internal target tumour. Three implanted FMs were used to collect the tumour motion for the CBCT projections, and an external marker was used to acquire the RPM signal and synchronise it with the motion.

2.4 Electromagnetic markers

This section describes the systems that can be used to monitor the motion with electromagnetic markers. The continuous 3D localization in real-time embedded transmitters/transponders is provided by electromagnetic systems eliminating the need for ionizing radiation. Calypso is a widely adopted non-ionizing medical device used for real-time tumour tracking and localization during RT. To provide continuous and precise 3D localization of implanted transponders, the system employs electromagnetic technology [112]. These transponders are implanted within or near the target tumour, allowing them to act as FMs. The system tracks the position of these transponders in real-time during radiation treatment and provides accurate information about the location and motion of the tumour. In this setup, a panel containing multiple excitation coils is placed above the patient. These excitation coils emit electromagnetic signals, which are used to stimulate the transponders one at a time. Each transponder responds by resonating with a unique electromagnetic frequency when excited. As the transponders are sequentially aroused, a second set of receiver coils, possibly placed around the patient or integrated into the treatment machine, detect the resonating signals emitted by the active transponder. The receiver coils pick up the electromagnetic responses from the transponders and relay the information to the tracking system. The tracking system then uses triangulation techniques, based on the time delay and intensity of the received signals, to accurately calculate the position of the resonating transponder in three-dimensional space. By triangulating the signals from multiple receiver coils, the system can precisely determine the transponder's location relative to the patient's anatomy.

The first clinical use-case was prostate cancer treatment from the Calypso [113]. One

of the key advantages of using Calypso in prostate cancer treatment was its ability to provide continuous monitoring without the use of ionizing radiation. This unique feature enabled Kupelian et al. [114] to conduct a systematic investigation of motion patterns in the prostate region. Shinohara et al. [115] evaluated the feasibility of implanting transponders with Calypso for intra- and interfraction motion monitoring in five pancreatic cancer patients. The mean shift from patient setup was used to assess interfraction motion in the X, Y, and Z axes, and the corresponding values were 4.5±1.0 mm, 6.4±1.9 mm, and 3.9±0.6 mm, respectively. The superior, inferior, left, right, anterior, and posterior mean intra-fraction motions were 7.2±0.9 mm, 11.9±0.9 mm, 2.2±0.4 mm, 3.1±0.6 mm, 4.9±0.5 mm, and 2.9±0.5 mm, respectively. The stability of the smooth transponder in lung tissue has posed a challenge as described in the study conducted by Shah et al. in 2013 [116]. To address this issue, an anchored version of the transponder with improved attachment within the bronchia has been developed. This modified version of the transponder includes five nitinol legs, which provide better fixation and stability within the lung tissue [117]. Later, Vanhanen et al. [118] conducted research on the potential impact of intra-fraction motion correction in prostate SABR by evaluating dose accumulation with Calypso-based continuous motion monitoring localization. The authors used 22 cancer patients with 308 fractions for this study. More recently, Capaldi et al. [119] developed a quality-assurance digital phantom for evaluating the performance of Calypso for lung cancers.

The Calypso system consists of certain limitations. One limitation is the detection range which may not extend adequately below the antenna panel limiting its ability to track targets located deeper within the body. Moreover, the system may lack flexibility when transferring between treatment rooms, necessitating a dedicated non-conducting

couch top, which could be cumbersome and time-consuming. Another concern is the potential for magnetic resonance (MR) artefacts caused by the transponders, which can affect the quality of MR imaging (MRI) in patients with implanted markers [120]. Additionally, the size of the transponders in the first-generation Calypso system was larger than standard FMs, potentially leading to more invasive implantation procedures. However, more recent developments have introduced thinner transponders that can be inserted with a 17-gauge needle to address this issue.

2.5 Ultrasound augmented monitoring

Ultrasound systems (US) with good soft-tissue contrast are capable of providing real-time continuous image acquisition, enabling clinicians to visualize internal structures dynamically. One significant advantage of ultrasound is that it does not involve ionizing radiation, making it a safer imaging modality for patients, particularly when repeated imaging sessions are required. One of the key strengths of ultrasound lies in its ability to directly observe the deformation of internal tissues in real-time. With high spatiotemporal resolutions, ultrasound can capture even subtle changes in tissue position and shape during various physiological processes, such as respiration or anatomical motion. Elekta's Clarity Autoscan is a commercial system designed for tracking intra-fraction motion [121]. This is used particularly to monitor the motion of the prostate during treatment delivery.

The US can track a number of anatomical surrogates as a modality of soft tissue imaging where it is difficult to distinguish the lesion. US-based methods widely used to monitor internal anatomical motions (including both intra- and interfraction) for prostate cancers [122, 123, 124]. It inspired the use of advanced USs to study a num-

ber of treatment sites beyond the prostate. Liver motion monitoring was assessed in a free-breathing patient immediately after liver SABR using an adapted Vivid 7 Dimension probe against Calypso [125]. For abdominal regions, an experimental analysis using US scanning has been conducted by several research groups by using breath-hold RT to track the liver's 3D position [126, 127]. Recently, Tianlong et al. [128] conducted research to track the intra-fraction tumour motion in the pancreas with USs using an abdominal phantom.

The careful positioning of the probe is required for optimal imaging to optimize patient interaction. Fargier et al. [129] and Li et al. [130] have identified a need for anatomical deformation control and changes in image quality associated with variations in the pressure of the probe. In addition to that, the probe must be manually calibrated during the patient setup to ensure both sufficient coverage of the target volume and reproducible positioning. To optimize the location of the probe during patient setup and radiation delivery, Sen et al. [131] developed robotic systems together with remote probe support. Further considerations are necessary when placing an ultrasound probe within the gantry arc due to potential implications on beam attenuation. This placement may affect the passage of the radiation beam, leading to alterations in dose delivery and potentially impacting treatment efficacy.

2.6 MRI for motion management

The use of MRI technology to aid RT has recently been implemented in clinical practice. The main reason to introduce this MR-guided RT is that it provides radiation-free, wonderful soft-tissue contrast and high-resolution images for RT treatment planning and motion management [10]. Using this approach, patients are not exposed to addi-

tional radiation and are not required to have FMs implanted within their body. With the advent of the MR-linac, some real-time imaging information (intersecting 2D planes, updated at 5-10 Hz) has recently become available. However, this technological solution is expensive and has a high associated time cost, limiting availability to a small cohort of patients in a few centres internationally. Additionally, metal-implanted cancer patients and/or very large patients could not be inspected with MR imaging. In this section, we describe the previous studies that have been done on MR imaging for real-time motion management.

MR-linac can acquire real-time imaging in two orthogonal scan planes and not yet possible to acquire, reconstruct and post-process 3D imaging at an acceptable resolution and imaging rate to estimate the motion of the target tumour. A single radio-frequency pulse is generally used in gradient-echo MR sequences, which are the foundation of cine MR imaging [132, 133]. This pulse helps to create a magnetic field gradient that is used to generate the images. It is feasible to generate a sequence of images that depict the movements of the tissues and fluids that are captured over time by using gradients of various strengths and orientations. These images can be used to create a movie-like sequence that provides a detailed view of the movement of the tissues and fluids being imaged.

By acquiring MR images from only a subset of k-space, the amount of data that needs to be processed is reduced, which can lead to faster image processing times. The process of gathering data from k-space, which is a mathematical representation of the spatial frequencies that make up an image, is referred to as under-sampling. A variety of techniques, such as randomly sampling k-space or collecting data solely from selected regions of k-space, can be used to achieve under-sampling. However, under-

sampling k-space can result in a loss of image detail and the introduction of artefacts, so careful optimization is necessary to balance processing speed with image quality. For example, parallel imaging methods may be used to reconstruct undersampled k-space data using several independent coils [134]. The additional spatial knowledge can be used during image reconstruction since the signal generated by each coil depends on its location relative to the patient. This allows for improved image quality and the ability to reconstruct images with higher spatial resolution. However, many MR-guided RT systems have limited parallel imaging capabilities compared to diagnostic MR scanners that are commercially available. This is a hardware limitation that needs to be addressed to improve the performance of MR-guided RT systems.

The location of a target tumour can be determined directly by tracking the tumour's motion over time using sequential MR images. However, these images can also be used indirectly by identifying a surrogate structure that correlates with the motion of the target tumour. For instance, if the tumour is in the lung, its movement can be linked to the diaphragm's movement. By analyzing the motion of anatomical structures using deformable motion models, it is possible to gain a more accurate understanding of how these structures move in 3D space. These models can be created by mapping 2D cine MR images to 3D anatomical models, which can then be deformed and transformed to match the motion of the structures observed in these 2D images [135, 136]. With the aid of segmentation or deformable image registration techniques and 2D cine MR images, a number of algorithms have been developed to efficiently and accurately extract the position or outline of a volume of interest [137, 138, 139, 140]. Even though 2D imaging methods, such as 2D cine MR, can provide high-resolution images of structures in a single plane, they may not provide enough information to accurately assess

the location and scale of the structure in 3D space. This is because structures in the body can move and deform in complex ways, and a single 2D image may not capture all the relevant information about the structure's 3D position and shape. Therefore, To improve the orientation of 2D cine MR images for adaptive RT in real-time, numerous studies have been conducted [122, 141, 142].

MR guidance for the monitoring of intra-fractional motion is still at its early stage. Using the ViewRay MRIdian, a few clinics have started to implement on-board MR imaging to control intra-fractional therapy beam gating [143, 144, 145]. For the first MR-guided procedure, which was used to treat tumours in thoracic and abdominal regions, around 33% of patients receiving MR-guided RT were diagnosed with gating [146]. Insight into the potential of using intra-fractional motion control for MR imaging can be gained from preliminary clinical trials and further research studies [147, 144]. To achieve this, special attention must be paid to the MR-linac's positioning and the use of specialized radiation delivery techniques, which can be challenging and require specialized training. Recently, Evan et al. [148] conducted a study using an MR-linac to evaluate the possibility of using continuous positive airway pressure, with or without DIBH, to control respiratory motion during treatment delivery with six healthy patients.

Uijtewaal et al. [149] proposed an MRI-guided multi-leaf collimator (MLC) tracking approach to monitor the tumour motion throughout intensity modulated radiation using an Elekta research tracking interface. The motion was generated using a Quasar MRI 4D phantom with and without 1.0 mm/min cranial-drift. The authors used a template matching method based on Cross-correlation to predict the positions of phantom tumours in sagittal 2D cine-MRI. They used two ways to train a linear regression model,

one based on several traces and the other based on a single trace, to optimize for online MRI and account for the expected system delay. Later, Subashi et al. [150] proposed a method that utilizes peripheral k-space view-sharing and a quasi-random projection-encoding sampling function to enhance the spatiotemporal resolution of respiratory motion in 4D-MRI. The authors optimised the spatial resolution and reduced temporal blurring effects of the MRI by directly extracting the respiratory signal from k-space without using any surrogate marker. More recently, Tallet et al. [151] used 59 liver cancer patients in a study to compare the use of MRI-linac with conventional IGRT for SABR. The authors reported that the boundaries of the liver tumours were not visible in any of the cases where CBCT was used as an IGRT tool, however when MRI was utilised as an IGRT tool, the tumour boundaries were evident in 72% of the cases.

2.7 Learning-based Techniques

2.7.1 Machine learning with respiratory motion tracking

Today, the widespread use of machine learning methods in the field of managing respiratory motion, as well as other medical applications, is growing rapidly. This section explores the role of machine learning in RT in the context of motion modelling. Supervised and unsupervised learning approaches are used to predict respiratory motion in short time intervals [152, 153]. In general, AI methods are used in image-guided 4D RT to take full advantage of the knowledge provided by radiographic verification and tumour tracking [154].

Recently, deep learning techniques have shown their remarkable performance and impressive learning power in analyzing numerous types of images including medical

images [155]. These approaches normally beat other different methodologies in the previously mentioned fields, which proves that deep learning can capture the semantic information of the data by learning robust features. Most of the recent work in real-time motion management has been used in deep learning techniques.

Several recent experiments have employed deep learning techniques to anticipate lung motions [156, 157, 158]. Some of these studies, such as [156, 158], have incorporated the concept of RNNs to construct a predictive model for pulmonary movements. The proposed methodologies aim to forecast the tumour's subsequent position based on its current location, dividing the data into a training set and a test set. RNNs, leveraging hidden neurons, memorize the relationships within input sequences as historical information, comprehending how elements transform and operate [159].

A study conducted by Kai et al. [156] utilized an RNN to predict lung motion, a technique subsequently applied in RT to model lung tumour trajectories. The motivation behind employing RNNs was to accurately estimate the future position of the tumour to compensate for the approximately one-second delay in the movement of the clinical linac gantry. Therefore, the study aimed to accurately estimate the tumour's position one second ahead, with a maximum allowable prediction error of 1mm in 3D space. To achieve this, the authors utilized three separate RNN models to estimate lung tumour trajectory for the x, y, and z axes motions for each patient case. They inputted past coordinates over a period of four seconds, with each data point in the measured tumour trajectory representing a sampling interval lasting 1/30 seconds. With 120 past coordinate data samples used for the estimation of the future position, the RNN's input layer consisted of 120 nodes to supply this data, while the hidden layer comprised 10 nodes. The RNN predictor was designed to forecast the tumour's position one second

ahead. They set a prediction horizon of 30 data points, approximately equivalent to one second into the future. The authors compared the proposed RNN model with a three-layer ANN formulated for a single-axis prediction for each patient case. The study calculated RMSE values of the predicted error in 3D space for all seven patient cases for both models. For the RNN model, these error values were 0.6822, 1.3720, 0.5957, 0.8612, 0.4799, 0.8160, and 0.9213 mm, whereas for the ANN model, they were 1.9684, 4.0334, 3.9098, 7.8485, 2.7157, 8.8041, and 8.4417 mm. The RNN predictor generated the predicted trajectory for six out of seven patient cases with an RMSE of less than 1 mm in 3D space. Another approach was introduced by Park et al. [157], who proposed a method based on intra-fraction and inter-fraction fuzzy deep learning. This technique not only predicted the breath-induced motion of the target tumour but also reduced computational time. The RMSE showed a noteworthy improvement of 29.98%. Later, Wang et al. [158] aimed to enhance the effectiveness of RT treatment during sessions by real-time prediction of tumour motion. To achieve this objective, the authors employed a bidirectional LSTM network. The dataset comprised respiratory motions from 103 patients with malignant lung tumours.

Steiner et al. [160] conducted a study to investigate whether measurements from both 4D-CT and 4D-CBCT images could effectively predict the range of motion of the target area during SABR treatment for lung cancer. In this research, Calypso beacons were implanted in 10 patients undergoing lung SABR. The null hypothesis posited that there would be no significant difference between the measurements obtained from 4D-CT and 4D-CBCT images and the range of motion of the target area during SABR treatment for lung cancer. Conversely, the alternative hypothesis suggested a significant difference in these measurements. The authors calculated the RMSE for each

phase by analyzing the reconstructed motion, imaging, and treatment motion. They explored the relationship of motion ranges in three directions: AP, LR, and SI views. Their findings revealed a rejection of the null hypothesis due to a Pearson correlation of less than 0.0001. They observed that both 4D-CT and 4D-CBCT significantly underpredicted the motion ranges of the treatment target during SABR treatment, with factors of SI=1.7, AP=1.7, LR=1.9, and SI=1.5, AP=1.6, LR=1.6, respectively.

Chenga et al. [161] presented a method employing RNN to predict heart motion using US images, aiming to enhance computational efficiency by mitigating the time associated with image acquisition and processing. The sequences of acquired US images were processed through an image processing algorithm to determine the position of interest using a surgical instrument captured by the same US scanner. Subsequently, the collected data points were input into an RNN to predict heart motion. Two types of datasets were utilized: the first comprised a fixed heart rate and maximum amplitude, while the second involved varying heart rates and maximum amplitudes. The authors assessed their approach using RMSE and mean absolute error (MAE), comparing the results with those obtained from an extended Kalman Filter (EKF) algorithm. Two neural network models were developed for the two datasets separately. In the first dataset, there was a 60% reduction in both MAE and RMSE compared to the EKF, while there was an approximate 70% reduction than the EKF when utilizing the second dataset for both evaluation metrics.

Lin et al. [162] devised a real-time respiratory signal prediction method based on deep learning, employing an LSTM model since the target respiratory motion needs to be predicted ahead of time a certain margin during the treatment delivery to accommodate for the latency associated with beam and field adjustments. A total of 1703 sets

of respiratory signals were gathered through an RPM system from 985 patients. The dataset was partitioned into training, internal validity, and test sets, with 1187 respiratory curves designated for the training set and the remaining 516 for the test dataset to ensure unbiased estimation of generalized performance. During training, each signal was split into training and internal validity parts. The LSTM model received input vectors containing 100 data points (corresponds to the length of time lag), representing segments of the breathing signal, with the aim of predicting the next 15 data points immediately following the input. This prediction task was achieved using a sliding window approach, where input and output pairs were extracted from the signal and moved along the time axis. The sliding window was shifted by 15 data points to continuously predict subsequent data points after the training input, and errors were calculated for each set of 15 predicted data points. The authors fine-tuned hyperparameters using an exhaustive grid search strategy and assessed the proposed LSTM model using three evaluation metrics: MAE, RMSE, and Maximum Error (ME). In the internal validity dataset, the LSTM model achieved 0.037, 0.048, and 1.687 for average MAE, RMSE, and ME, respectively. For the test dataset, the corresponding values for these evaluation metrics were 0.112, 0.139, and 1.811. However, this model is not suitable for real-time target motion monitoring, as it can only predict external respiratory signals.

Teo et al. [163] proposed a method to predict tumour motion during treatment delivery, employing a multi-layer perceptron (MLP) network trained through a combination of online and offline learning on tumour trajectories. The final model comprises a single hidden layer MLP with 20 neurons. The study utilized 35 input data samples, with an average sliding window size of 28 data samples. Hyperparameters were fine-tuned through a trial-and-error approach, and the model's performance was assessed

using two evaluation metrics: MAE and RMSE. To evaluate the generalized model performance, 20 tumour traces were used in a leave-one-out cross-validation. The MLP model demonstrated an overall MAE of 0.57±0.17 mm and an average RMSE of 0.67±0.36 mm.

Huang et al. [164] developed a deep learning approach for motion-compensated dynamic MRI reconstruction to enhance image quality using under-sampled MRI k-space data. This complex problem involves three key tasks: dynamic reconstruction, motion estimation, and motion compensation. For model-based dynamic reconstruction, the authors employed an RNN-based technique, specifically a convolutional gated recurrent unit (ConvGRU) architecture with a U-Net serving as an encoder and two decoders. U-Net acted as the backbone, while ConvGRU detected the dynamic behaviour of the image sequence. In the second component, they utilized a CNN-based architecture called U-FlowNet to estimate the motion field. In the third component, the estimated motion was applied to the reconstructed images to refine and generate a motion-compensated image. The authors utilized a short-axis cardiac dataset from fifteen patients, employing a 3-fold cross-validation technique for evaluation. Quantitative metrics such as RMSE, peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM) were employed. Later, the authors published an extended version of this approach to further enhance the quality of the reconstructed results [165].

Mafi et al. [166] devised a real-time respiratory motion prediction approach using a neural network. This design incorporated a three-layer static feed-forward ANN connected to a single-layer RNN in the second hidden layer of the ANN. The network was trained on a dataset of tumour motion, comprising 143 treatment fractions from 42

thoracic and abdominal cancer patients. The dataset was partitioned into sets of 100, 13, and 30 signals for training, validation, and testing, respectively. A fixed input sliding window size of 35 data samples was utilized, along with a five-sample prediction horizon to minimize system latency. The model aimed to capture the fifth oncoming sample. The authors conducted various experiments, altering the output window size to 1, 3, and 5, and assessed the results using RMSE for three types of models: static ANN, static ANN + online retraining, and the proposed dynamic neural network. The findings indicated that the dynamic neural network model with an output window size of 3 achieved the lowest RMSE, establishing itself as the optimal model for motion prediction.

Lee et al. [167] introduced a method to enhance image quality by mitigating streaking artefacts arising from sparse-angle projections in 4D CBCT. The authors used a convolutional neural network (CNN), in particular, residual U-Net with a wavelet-based approach. The proposed approach was compared against three existing methods: filtered-back-projection (FBP), compressed sensing (CS), and a simple residual CNN. Image quality was assessed using RMSE, universal quality index (UQI), and SSIM. The proposed approach yielded RMSE values 0.24, 0.22, and 0.017 times lower compared to CS, FBP (4D), and FBP (3D) algorithms, respectively. Correspondingly, SSIM and UQI values for the proposed technique were 0.950 and 0.998.

Lin et al. [168] proposed an ensemble method for estimating the motion of lung tumours in 3D space along the AP, SI, and LR directions, employing machine learning algorithms. The proposed model architecture incorporates four foundational machine learning models: XGBoost, LightGBM, MLP, and random forest. The training utilized 16 visual features derived from non-4D CT images and 11 clinical features extracted

from the Electronic Health Record (EHR) database encompassing 150 patients. The assessment was conducted based on MAE and RMSE. In the SI direction, the corresponding values were 1.23 mm and 1.70 mm, in the AP direction, the predictions were 0.81 mm and 1.19 mm, and for the LR direction, the error values were 0.70 mm and 0.95 mm.

Zhang et al. [169] introduced a deep learning-based technique for 4D-CBCT image reconstruction that incorporates simultaneous motion estimation and image reconstruction. This method was evaluated with a focus on fine details in lung 4D-CBCTs. The model architecture, based on U-Net, was employed for estimating the deformation vector field (DVF), which represents the final position of each voxel after it has been transformed, i.e. it is the initial position of the voxel + the displacement of the voxel, to enhance the accuracy of intra-lung DVFs. Subsequently, Zhehao et al. [170] proposed a CNN-based approach to enhance image quality by reducing motion artifacts in phase-correlated Feldkamp, Davis and Kress (FDK) reconstructed 4D-CBCT images. To gauge the effectiveness of their method, the authors utilized the XCAT phantom and SPARE dataset, quantitatively assessing the results using RMSE and normalized cross-correlation (NCC). The outcomes revealed that the authors achieved an RMSE error of 0.0021 ± 0.0003 and an NCC value of 0.93 ± 0.04.

Mendizbal et al. [171] presented a method employing a U-Net variant architecture capable of learning the function describing the relationship between an input force and the resulting deformation across diverse geometries, enabling rapid predictions. The authors utilized finite element (FE) simulations to acquire deformations for the training process. When subjected to an applied force, the U-Net demonstrated the ability to approximate deformations in the liver anatomy, achieving an MAE of 0.22 mm with

a prediction time of 3 ms. The model requires, as input, point clouds derived from tetrahedral meshes mapped to a sparse hexahedral grid. This study assumes uniform material characteristics throughout the deformable region.

Recently, Shi et al. [172] introduced a temporal CNN-based approach for estimating respiratory motion in thoracoabdominal tumours. The study utilized a dataset comprising 103 cancer patients to generate the motion dataset. In related work, Li et al. [173] employed a machine learning-based method that leverages radiomic features extracted from average intensity projections to estimate motion in the lung (using 33 radiomic features) and liver tumours (using 22 radiomic features). The dataset for this study included 108 lung and 71 liver cancer patients. Validation involved 26 independent models, 13 for lung motion, and the remaining for liver motion. The achieved maximum sensitivity and specificity for lung motions were 0.848 and 0.936, respectively, while for liver motions, they were 0.862 and 0.829.

2.7.2 Image registration approaches for motion estimation

Lv et al. [174] presented an approach based on employing a CNN within a deformable image registration (DIR) technique to estimate deformations caused by respiration in the abdominal region. The study utilized MR images obtained from ten patients through a 1.5T MRI system. Employing non-Cartesian iterative SENSE reconstruction, the acquired images were organized into three bins based on corresponding respiratory signals. Subsequently, the authors utilized a trained CNN model to assess the spatial transformations among the bins. Comparative evaluations with local affine registration and non-motion corrected registrations demonstrated better registration results.

Sokooti et al. [175] introduced a method centred on non-rigid image registration, employing a multi-scale CNN architecture to estimate the DVF. The study utilized 3D CT chest data and conducted intra-subject registration by applying the estimated deformation to the images. In contrast to early fusion, this method opted for late fusion, where patches are concatenated and utilized as inputs for the network. The system demonstrated competitive performance compared to B-Spline registration.

Eppenhof et al. [176] proposed a 3D CNN based on U-Net architecture designed for deformable image registration of Pulmonary CT images with synthetic random transformations. The DIRLAB dataset was employed, revealing an average Target Registration Error (TRE) value of 2.17±1.89 mm. In comparison to actual lung motion, random transformations exhibited significant differences. Consequently, supervised training with random transformations failed to provide accurate regularization of the DVF. To conserve memory and prevent loss of image information, the authors suggested the use of a downsampled image rather than the entire image during the training phase.

Uzunova et al. [177] applied a CNN for estimating the deformation region in the context of 2D brain MR and 2D cardiac MR registration. They also adapted FlowNet for dense image registration. The study involved generating numerous synthetic image pairs with corresponding ground-truth deformations by learning a statistical appearance model from a limited number of sample images. The pre-trained FlowNet architecture was fine-tuned using these synthetic data, and the results demonstrated that the data-driven, model-based augmentation approach outperformed generic but highly unspecific methods.

Giger et al. [178] presented a patient-specific motion modelling approach using a con-

ditional Generative Adversarial Network (GAN)-based image registration technique. The model underwent training to understand the relationship between navigator-based MRI images and their corresponding US images. By utilizing US images as surrogate signals, it effectively anticipated 3D-MRI image volumes associated with different respiratory states. The evaluation of this methodology included three lung cancer patients. However, a significant drawback of this method is its vulnerability to adverse effects in instances of slight displacements of the US probe.

Fu et al. [179] introduced an approach reliant on unsupervised deformable registration for estimating the DVF associated with lung motion. This method employs two GAN sub-networks, namely CoarseNet and FineNet. Initially, CoarseNet performs whole-image registration on a down-sampled image, followed by the use of the patch-based FineNet to register image patches from the globally warped moving image to those of the fixed image. The study utilized ten 4D CT images with five-fold cross-validation, and an additional ten datasets from the DIRLAB [180] data repository were employed for comprehensive comparison. To enhance registration accuracy, vessel enhancement was applied before DIR. TRE was used as the evaluation metric, with an average TRE of 1.00±0.53 mm for their dataset and 1.59±1.58 mm for the DIRLAB datasets.

Sokooti et al. [181] presented a supervised approach for non-rigid image registration, generating ground-truth DVF for model training. During the training phase, they utilized randomly generated transformations with both single and mixed frequencies. The study involved a comparison of different network architectures, such as U-Net applied to the entire image and an advanced U-Net applied to image patches. Performance evaluations were conducted based on the TRE and Jacobian determinants. The experiment indicated that the network trained with model-based respiratory motion

outperformed networks trained with random transformations.

Sentker et al. [182] proposed a deep learning-based framework for rapid 4D CT image registration. In their approach, 4D CT images obtained in-house were utilized for training, while external evaluation cohorts comprised open 4D CT data repositories. They employed uncertainty maps based on dropout for analyzing different variations of the proposed framework. By comparing their framework to standard DIR, they demonstrated that the registration accuracy is comparable, with a speed-up factor from around 15 minutes to a few seconds (speedup of approximately 60-fold).

Qin et al. [183] conducted a study employing a biomechanics-informed neural network for image registration, specifically focusing on myocardial motion tracking in 2D stacks of cardiac MRI data. The architecture utilized a variational autoencoder (VAE) to learn a manifold for biomechanically plausible deformations, based on reconstructed biomechanically simulated deformations. The trained VAE was then integrated with a deep learning-based image registration network, providing a parameterized registration function that was regularized by application-specific prior knowledge to generate biomechanically plausible deformations. Although their method demonstrated superior performance compared to reference methods, it is currently limited to 2D motion tracking.

Zhang et al. [184] developed an approach for estimating high-quality CBCT image volumes with limited-angle on-board kV X-ray projections, employing an unsupervised 2D-to-3D deep learning-based (U-net) deformable registration technique. The model's inputs consisted of a high-quality CBCT image volume (source image) and a reconstructed 3D-CBCT (target image), generated from a series of highly sparse kV images acquired with limited angles. The model produced a predicted dense DVF.

During the training process, the authors integrated a non-trainable Siddon-Jacobs ray tracing algorithm [185, 186] to generate DRRs from the deformed 3D-CBCT image volume. Forward similarity loss was calculated between these projections and the target kV X-ray projections used to generate the reconstructed 3D-CBCT. Additionally, the authors incorporated inverse similarity loss between DRRs from the input and the predicted high-quality source image into the total loss function. The predicted high-quality source image was generated by applying the spatial transformation from the inverse DVF to the deformed high-quality 3D-CBCT. Subsequently, the same authors [187] applied a similar technique to reconstruct high-resolution MRI with limited k-space data, enabling real-time tracking of anatomical motion.

Lee et al. [188] proposed a patient-specific registration framework based on deep learning to estimate the rigid transform parameters of C-arm pose from intraoperative fluoroscopy images. The authors generated training data using a parameterized breathing motion model derived from patient-specific pre-operative 4D CT data. The method underwent evaluation on both a synthetic test dataset and real preclinical swine fluoroscopy images, with assessments made using 3D mean-target-registration-error (mTRE) and mean-projection-distance (mPD). Results for the synthetic test dataset showed an mTRE error of 6.4±3.3 mm and an mPD error of 7.8±3.9 mm. For real fluoroscopy images, the mPD error was 14.1±2.7 mm. Subsequently, Lecomte et al. [189] conducted a study aiming to estimate a dense 3D displacement vector field from a single fluoroscopy image, utilizing a 2D-to-3D deep learning-based non-rigid registration technique. The model underwent evaluation on 4D-CT lung data, achieving a landmark error range of 2.3 to 5.5 mm with mTRE.

Recently, Xie et al. [190] presented an unsupervised deformable registration method

for inter-fraction CBCT-CBCT images based on deep learning. The entire network comprises a global and a local GAN to estimate coarse-to-fine level deformation fields. The total loss function integrates three components: a similarity loss, an adversarial loss, and regularization based on the DVFs. The model underwent training on 100 fractional CBCTs using five-fold cross-validation and evaluation on an additional 105 CBCTs from 20 and 21 abdominal cancer patients. The authors reported an average TRE of 1.91±1.18 mm, computed based on landmarks and implanted FMs.

Later, Dong et al. [191] proposed a 2D-3D non-rigid registration technique for monitoring lung tumour motion using deep learning. This approach incorporates two orthogonal DRR projections to predict the 3D DVF. Initially, 3D feature maps were extracted from the orthogonal DRRs using a series of residual blocks. These feature maps were then utilized as the fixed image, with a reference 3D-CT serving as the moving image input to an attention-based U-Net architecture. The registration process was completed in 1.2 seconds, yielding a dice coefficient exceeding 0.97 and a normalized cross-correlation surpassing 0.92.

In a very recent study, Dai et al. [192] introduced a patient-specific 2D-3D deep learning-based image registration method to monitor volumetric lung tumours using a single kV projection image. Initially, synthetic motion instances, i.e., deformed 3D-CTs and corresponding segmentations, were generated from 4D-CT data through a hybrid data augmentation technique. From the augmented data, 9000 samples were assigned for model training, while 500 samples were set aside for validation and another 500 for testing. To align the DRR images with real kV images, a Contrastive Unpaired Translation GAN model [193] was employed to transfer the style. The model was trained to predict 3D deformation fields utilizing a spatial transformer network to

transform the planning CT volume and the corresponding segmentation mask volume. The study utilized a 4D-CT patient dataset from the TCIA image archive [194, 195, 196, 197], as well as 4D-CT data and 2D real kV images from CIRS phantom data. The model's performance was evaluated using real kV images from the CIRS phantom data, specifically at projection angles of 0 and 90 degrees. The results demonstrated that the authors achieved an RMSE of less than 1.5 mm for the tumour centroid when compared to real kV images.

2.7.3 Surrogate-driven motion models

Surrogate-driven motion models estimate internal anatomical motion using some surrogate signal under the assumption that the two are well correlated. External signals, such as markers on the skin surface, or internal signals, such as diaphragm motion, can be used as surrogate signals. A detailed review of existing respiratory motion models was published by McClelland et al. [198]. Later, McClelland et al. [199] proposed a technique that integrated the processes of motion model creation and image reconstruction, using partial imaging data (i.e. slab and slice images) and unsorted 4D-CT data as input, together with a corresponding surrogate signal. Image registration was used to extract internal motion from dynamic images, which was then used to fit a correspondence model that related these motions to surrogate signals. This model has two degrees of freedom as the authors fit the model using two surrogate signals, and therefore can simulate variable motion including intra- and inter-breath variability. They evaluated their approach with 4D-CT data using two sets of manually annotated landmark points and 2D phantom data using displacement field error. Mean errors of 1.88 mm and 1.72 mm were reported for two landmark sets. Meanwhile, Guo et al. [200]

presented a Motion-Compensated Simultaneous Algebraic Reconstruction Technique (MC-SART) which is capable of reconstructing high-quality images and motion models from CBCT projections and respiratory surrogate data. These techniques [199, 200] can be used to estimate an image volume at any given time point using surrogate measurements based on a volumetric reference image with an estimated motion model.

For MRI-guided RT, Stemkens et al. [135] and Harris et al. [201] constructed models that inferred 3D motion from 2D cine-MR images obtained using a 2D image navigator. The former group [135] tested their method on seven healthy volunteers and a torso-shaped, MRI-compatible motion phantom for pancreas and kidney. Harris et al. [201] assessed their method using both digital extended-cardiac torso (XCAT) lung cancer simulations and MRI data from four liver cancer patients. Tran et al. [202] analyzed various MRI-derived surrogate signals to predict internal anatomy respiratory motion, including breath-to-breath variability and sliding motion. The models were evaluated on eight lung cancer patients by estimating the 2D motion from coronal and sagittal slices. Mean errors for coronal and sagittal slices were around 1 mm and 0.8 mm, respectively.

Some very recent efforts have focused on deep learning-based models. Romaguera et al. [203] developed a conditional-GAN-based probabilistic model which relies on in-room surrogate data. 3D volumes are estimated using a pre-operative static volume and 2D surrogate images. They evaluated their approach on 25 healthy volunteers and 11 cancer patients with free-breathing 4D MRI and ultrasound imaging datasets. They achieved a mean error of 1.67±1.68 mm for volumetric prediction from surrogate images, and 2.17±0.82 mm for unseen patient US and MRI cases. Moreover, using the MRI dataset, they achieved a mean landmark error of 1.4±1.1 mm. Using

convolutional auto-encoder and 2D surrogate ultrasound images, recently Mezheritsky et al. [204] proposed a surrogate-driven deep learning technique for population-based respiratory motion modelling. To execute inference, two pre-treatment 3D volumes of the liver at extreme breathing phases are required, as well as live 2D surrogate images reflecting the organ's current state. The model was evaluated using 4D ultrasound images from 20 volunteers with a reported mean tracking error of 3.5±2.4 mm. Liu et al. [205] proposed a neural-network-based approach to representing lung motion by predicting 3D CT volume at a given time point using diverse surrogate signals. The authors used a thorax phantom and seven lung cancer patients implanted with FMs to conduct their study. The authors achieved an average error of 0.66 mm for marker localization.

Internal motion estimates based on easily measured external surrogate signals are a potential means of acquiring requisite information on tumour position. However, these can be inaccurate depending on the strength of the correlation between tumour motion and surrogate; ambiguities in the displacement and phase relationships between the two may also be present. Moreover, such approaches are practically limited to describing aperiodic motions, since acquiring aperiodic surrogates, with their temporal irregularity, is extremely challenging. Aperiodic motions, such as changes in anatomical structure, bladder filling status, or bowel motions generally are not considered at all, though they can result in considerable anatomical deformations. While we too focus here on respiratory motion, our approach involves no assumption of periodicity or respiration-specific motion patterns. Given the means of generating requisite training data, it could be applied to motion of any sort.

2.7.4 3D Shape reconstruction from single-view projections

This section describes a class of methods that, more closely matched with our proposed approach, aims to reconstruct 3D geometry from 2D images. Several reports have described techniques for reconstruction from RGB images. For example, Wang et al. [206] proposed the GNN-based Pixel2Mesh algorithm, which deforms an ellipsoidal surface mesh using CNN-derived semantic characteristics from an input image, and applied it to an analysis of natural shapes (aeroplanes, chairs, cars, etc.). The ellipsoidal starting mesh limits the approach to genus-0 shapes, though in principle it could be adapted to other topologies. Smith et al. [207] extended the method to better capture local surface geometry, though the topological constraints remained. Similar ideas were used in [208, 209] to reconstruct 3D human body shapes from single RGB images. In the medical domain, Wu et al. [210] proposed a CNN architecture for reconstructing 3D lung shapes, in the form of point clouds, from a single-view 2D laparoscopic image. The authors focused on reconstructing the 3D point-cloud from 2D colour images of organ surfaces obtained from laparoscopic video feeds, rather than 2D slices or projections through a volume, which is a completely different problem than ours.

While clearly sharing elements of our target problem, reconstruction from RGB images rather than 2D projections is nonetheless a substantially different one. Various approaches addressing the latter scenario have appeared recently. Ying et al. [211] proposed X2CT-GAN to reconstruct 3D-CT volumes from bi-planar 2D X-ray images using GANs. The authors used 1018 chest CT images, to generate paired sets of simulated X-rays and 3D-CT images. These images varied in capture ranges and resolutions, necessitating initial resampling to a uniform voxel size of $1mm \times 1mm \times 1mm$.

Subsequently, a cubic area of $320mm \times 320mm \times 320mm$ was cropped from each scan image. For the training dataset, 916 CTs were randomly selected, while 102 CT images were designated for testing purposes. Each CT image in the training and test datasets was used to generate two DRR images PA and lateral views (projection angles 0 and 90 degrees, respectively) with full FOV and a size of 128 x 128 pixels. The researchers evaluated the predicted CT images against ground-truth CTs using PSNR and SSIM metrics, yielding values of 26.19 ± 0.13 and 0.656 ± 0.008, respectively. GANs may struggle to accurately reproduce fine details and anatomical structures present in ground-truth CT scans. The synthetic nature of GAN-generated images can result in reconstructed artefacts or inconsistencies, such as smooth regions, distortions, or unrealistic textures, which can compromise the clinical accuracy and reliability of the reconstructed volumes [211]. The variability in anatomical motion during biplanar image acquisition in clinical settings can lead to inaccuracies or inconsistencies in the reconstructed 3D volumes. By this means, biplanar X-rays often capture the anatomy in different motion states, particularly during free-breathing scenarios. This variability can result in discrepancies between the captured images, making it difficult to align and integrate the information effectively for 3D reconstruction.

Tumour localization with a single-view projection approach was proposed by Wei et al. [212]. The authors first developed a principal component analysis (PCA)-based breathing motion model using planning 4D-CT data. Consequently, they generated 1000 3D-CTs with varied tumour positions by randomly sampling the PCA coefficients. A CNN model was then employed to predict these PCA coefficients based on input DRR images. The authors used an angle-dependent ROI 2D projection mask to remove pixels unrelated to respiration and a projection angle-dependent fully-connected

(FC) layer. This layer was designed only to handle the discrete level of angles ranging from 0 to 360, and only one group of weights and biases were used to generate the output of this layer for each degree of the projection angle. Due to this limitation, the authors chose a binary projection mask at the nearest integer even though it is a fractionated gantry acquisition during both training and application/test stages. The method was evaluated using 15 patient datasets, where data augmentation techniques were applied to address intensity differences between DRR and CBCT images, resulting in 10 augmented projections per DRR. The mean tumour localization error was measured under $1.8 \pm 0.6mm$ (SI direction) and $1.0 \pm 0.5mm$ (lateral direction) for visible tumour cases in projection images. For cases where tumours were not visible in projections, the mean localization error did not exceed $1.5 \pm 0.9mm$ in both directions. However, challenges appeared in cases with significant intensity variations between DRRs and CBCT projections, affecting localization accuracy due to intensity shift issues. Additionally, reconstruction artefacts in 4D-CT images, such as structural blurriness or duplications, influenced localization accuracy. The study suggests potential limitations in handling variations in breathing amplitude and patient setup during treatment, recommending retraining PCA and CNN models with repeated 4D-CT data acquisition for validation. Furthermore, tumour localization using binary projection masks proved challenging for certain cases with specific projection angles, as these projections contained little information related to breathing motion, making tumour position deduction difficult [212].

Wang et al. [213] also proposed a CNN-based approach for reconstructing lung surface shapes from single-view 2D projections. The authors utilized 4D-NCAT and 4D-XCAT digital phantoms to create a lung motion dataset, comprising 542 pairs of left

and right lung meshes along with corresponding deformed 3D-CTs. Various shape deformations and spatial transformations were applied to simulate real lung shape variations, and resulting deformed 3D-CTs were used to generate full FOV DRR projections from front views. The training dataset included 446 pairs with corresponding DRR images, while the test dataset comprised 96 pairs. MobileNet [214] architecture, followed by a 1×1 convolution layer, was employed to extract image features. Subsequently, an FC layer was used to learn deformation parameters for mesh template control points, followed by another FC layer to adjust translation. The evaluation was based on metrics including Chamfer distance (CD), Earth mover's distance (EMD), F-score, and Intersection over union (IoU), yielding values of approximately $1.7mm$ for CD and $57-60mm$ for EMD, along with F-score and IoU scores ranging from 0.72 to 0.84 for left and right lungs, respectively.

Furthermore, the authors [213] assessed this approach's robustness using phase zero 3D-CT volumes from ten 4D-CTs in the DIR-LAB dataset, generating DRR images and corresponding left and right meshes for qualitative evaluation. Several limitations were identified in the study. Firstly, the approach was not evaluated under limited FOV settings, which are essential in clinical DRR/X-ray acquisition scenarios. Additionally, the evaluation lacked real X-ray images, despite utilizing real patient data from the DIR-LAB dataset. Another limitation was the exclusive use of front-view projections (i.e., projection angle zero) for evaluation, neglecting the challenges associated with different projection angles encountered in clinical practice. These limitations underscore the need for comprehensive evaluations under varying clinical conditions to assess the approach's practical applicability and performance in real-world scenarios.

Tong et al. [215] proposed X-ray2Shape to reconstruct 3D liver surface meshes by

combining GNN and CNN networks (the latter to encode image features). A mean shape derived from 124 patients was used as prior (i.e. initial template) and deformed by the GNN to match the individual organ shape. The authors employed a method where each vertex in the initial template mesh corresponding to a fixed angle (i.e. zero) was projected onto the front-view DRR image plane to derive pixel coordinates. These pixel coordinates were then utilized to extract relevant features from latent convolutional layers within a trained CNN encoder (VGG-16 [216]), which were associated with each vertex. The extracted features were concatenated with the corresponding 3D coordinates to generate a feature vector for each vertex. Subsequently, these feature vectors were incorporated into a GCN comprising eight sequential GCN layers to compute the deformation from the initial template to individual organ shapes. The difference between the estimated shape and the ground-truth shape was evaluated using mean distance metrics, including the mean value of the nearest bidirectional point-to-surface distance and the mean Euclidean distance, yielding results of approximately 6.71 mm and 16 mm, respectively. Later, the same authors [217, 218] extended the approach to reconstruct multiple abdominal organ shapes from a single projection image. These latter approaches [215, 217, 218] are designed to operate on images acquired at fixed projection angles—front view projections, equivalent to gantry angle 0 in our case—and therefore cannot directly accommodate images from arbitrary angles, as required here. Moreover, they predict only the organ surface shapes, rather than their full volumes.

Lu et al. [219] developed a CNN-based supervised learning approach to estimate the 3D-CT from a single DRR image. A 2D to 3D encoder-decoder architecture was used to first estimate the low-resolution 3D-CT followed by a super-resolution module based

on sub-pixel layers to reconstruct the high-quality 3D-CT. The authors used 4D-CT datasets (with 10 phases) acquired in three different fractions of a lung cancer patient. They used the first six phases for training, and the remaining four phases were equally divided into validation and test sets. They obtained a PSNR of 18.621±1.228 dB and an SSIM similarity score of 0.872±0.041. For reconstructing the 3D-CTs, the authors used only the front-view (i.e. projection angle zero) DRRs. The utilization of DRR images, while reasonable for simulation purposes, deviates from the real X-ray images acquired in clinical settings [219]. This discrepancy could impact the generalizability and reliability of the proposed approach. Additionally, the evaluation did not account for limited FOV settings, an essential scenario in clinical practice that could influence the method's performance and effectiveness.

A true volumetric approach for estimating liver deformations was proposed very recently by Shao et al [220]. Their method first used a GNN to predict the deformed liver surface. These deformations were then passed as boundary conditions to a finite element model of the liver, which computed the corresponding volumetric deformations. In this way, some level of biomechanical constraint was also introduced. The approach was evaluated for several projection angles (specifically: 0°, 45°, and 90°), however, the model required retraining for each angle; that is, each new angle effectively required a separate model. The model also required a very high number (3840) of image features to be encoded on graph nodes; in our approach, we use only 20. Finally, while biomechanical constraints can in principle be attractive for enforcing physical plausibility, the finite element solutions were in practice time consuming, which may be significant for clinical use, especially in-treatment adaption of therapy. Later, the same authors [221] extended this approach by incorporating an optical surface image

through a deep learning approach to estimate the motion of the liver boundary. Since projection images are usually acquired with a small FOV, the authors utilized optical imaging to incorporate a larger FOV of the body surface and then obtained motion correlation with the liver anatomy using a deep learning approach. This motion is then further corrected by using a GNN with a single kV planar x-ray projection, followed by a U-Net-based biomechanical modelling for intra-liver motion correction.

Our approach requires training only once and is applicable across all projection angles, unlike prior approaches cited in [215, 217, 218, 220, 221] that necessitate separate retraining for each angle. This limitation restricts their practical use in scenarios where the gantry rotates during beam delivery. Although these methods claim to be end-to-end trainable, they rely on a fixed, non-trainable projection step to extract semantic features, tailored specifically to a predetermined projection angle. Therefore, if they are to be used with all projection angles, they must be retrained separately for each unique angle. Additionally, all of these approaches utilized DRR images with a wide FOV to encompass the entire liver anatomy in their experiments. However, in clinical settings, projection images are typically acquired with a limited FOV, posing challenges in feature extraction by projecting vertices onto the DRR plane, as the projected surface mesh nodes may extend beyond the projection FOV, hindering feature extraction [220, 221].

2.8 Summary

The primary objective of this chapter was to explore the extensive literature review on motion modelling in RT conducted over the past few decades. The initial focus was on methods applicable to conventional RT linacs equipped with kV and MV X-

rays/CBCTs. The chapter then delved into approaches, including surface-guided methods, hybrid techniques, electromagnetic markers, and ultrasound-based strategies. An in-depth analysis was conducted to understand the advantages and disadvantages of these methods, with a common practice involving the use of FMs surgically implanted near the treatment target. However, FM-based approaches introduced potential issues such as organ inflammation, displacement, or migration during treatment delivery. In contrast, internal anatomical surrogates such as chest walls, diaphragms, or external markers/surrogates may give a poor correlation of the degree of motion between the target tumours and the surrogates and hence produce inaccurate real-time motion estimation.

Furthermore, a detailed review of the use of MRI for motion management was presented, highlighting its cost and time constraints, limiting its availability to a small cohort of patients in a few international centres. The majority of patients are currently treated on conventional linacs equipped with kV/CBCT.

The subsequent sections provided a comprehensive review of ML techniques, focusing on deep learning models, image registration techniques, and surrogate-based motion models, outlining their respective limitations.

Towards the end of the chapter, a detailed review of shape reconstruction from single-view projections was conducted. All of these approaches utilized DRR images with a wide FOV to include the entire liver anatomy for their experiments. However, in clinical settings, projection images are usually acquired with a limited FOV. Consequently, these approaches underwent training and validation solely on DRR images, without undergoing testing with real kV planar X-ray images. Moreover, most approaches utilized surface meshes, emphasizing the importance of reconstructing a 3D organ model

with internal density rather than solely focusing on surface texturing. Current methods for 3D organ reconstruction from single-view projections were acknowledged to be validated only for fixed gantry angles, limiting their applicability when the gantry rotates during beam delivery.

There is therefore a critical need to estimate the true motion of internal anatomy using single-view kV planar imaging, irrespective of gantry angles. This approach aims to provide continuous, genuine motion insights for a given organ through snapshots acquired at various time scales, without relying on invasive FMs.

However, before delving into the experiments, a detailed introduction to the generation of synthetic motion datasets for the experiments is required. The following chapter will provide a detailed description of this process, including preprocessing steps and solutions to problems encountered.

Chapter 3

Deep-Motion-Net: GNN-based volumetric organ shape reconstruction from single-view 2D projections

3.1 Introduction

This chapter describes a method termed Deep-Motion-Net, designed to facilitate the reconstruction of 3D (volumetric) organ shapes using a single in-treatment kV planar X-ray image obtained from any arbitrary projection angle. The primary aim of this chapter is to confront the challenges associated with patient motion in the context of radiotherapy, as outlined in Chapter 1. This is achieved through the implementation of an end-to-end deep learning architecture that learns the complex relationship between 3D anatomical motion and the corresponding anatomical appearance in kV images. The ultimate goal is to predict motion directly from such images, eliminating the need

for additional post-processing steps or invasive FMs.

The model learns a mapping from kV image-derived features to displacements of nodes in a patient-specific template organ mesh. Features are extracted by a CNN image encoder, while regression of the features with node displacements is learned by a GNN network. Importantly, the complete model is end-to-end trainable by virtue of a series of feature pooling networks (FPNs) that fuse image features with the 3D graph nodes, eliminating non-trainable components (i.e. vertex projection onto the 2D image space) that would otherwise be required. Finally, the model also learns projection angle-dependent features by encoding the angle in an additional channel to the input image. By this means, the model can reconstruct the 3D anatomy from kV images acquired at any projection angle. To the best of our knowledge, this is the first framework capable of reconstructing 3D anatomy from such inputs. While the method is general, this work focuses on respiratory motion and evaluates the method using synthetic and real images from liver cancer patients. A high-level overview of the workflow of our proposed method is illustrated in Figure 3.1.

The model underwent training using synthetic motion data, as detailed in Section 3.6, where ground-truth motions were inherently known through the construction of the data. The findings of this chapter ensure that the model is equipped with a robust foundation for understanding and predicting anatomical motion, offering a solution to the challenges posed by patient motion in the field of radiotherapy.

In Section 3.7, we delve into the overall methodology employed in our deep learning model. The evaluation of the model and the presentation of results are covered in Section 3.8, while the last section provides an in-depth description of the conducted ablation study.

This chapter is derived from my manuscript titled ”Deep-Motion-Net: GNN-based volumetric organ shape reconstruction from single-view 2D projections,” which has been submitted to the Medical Image Analysis journal and is presently undergoing the review process.

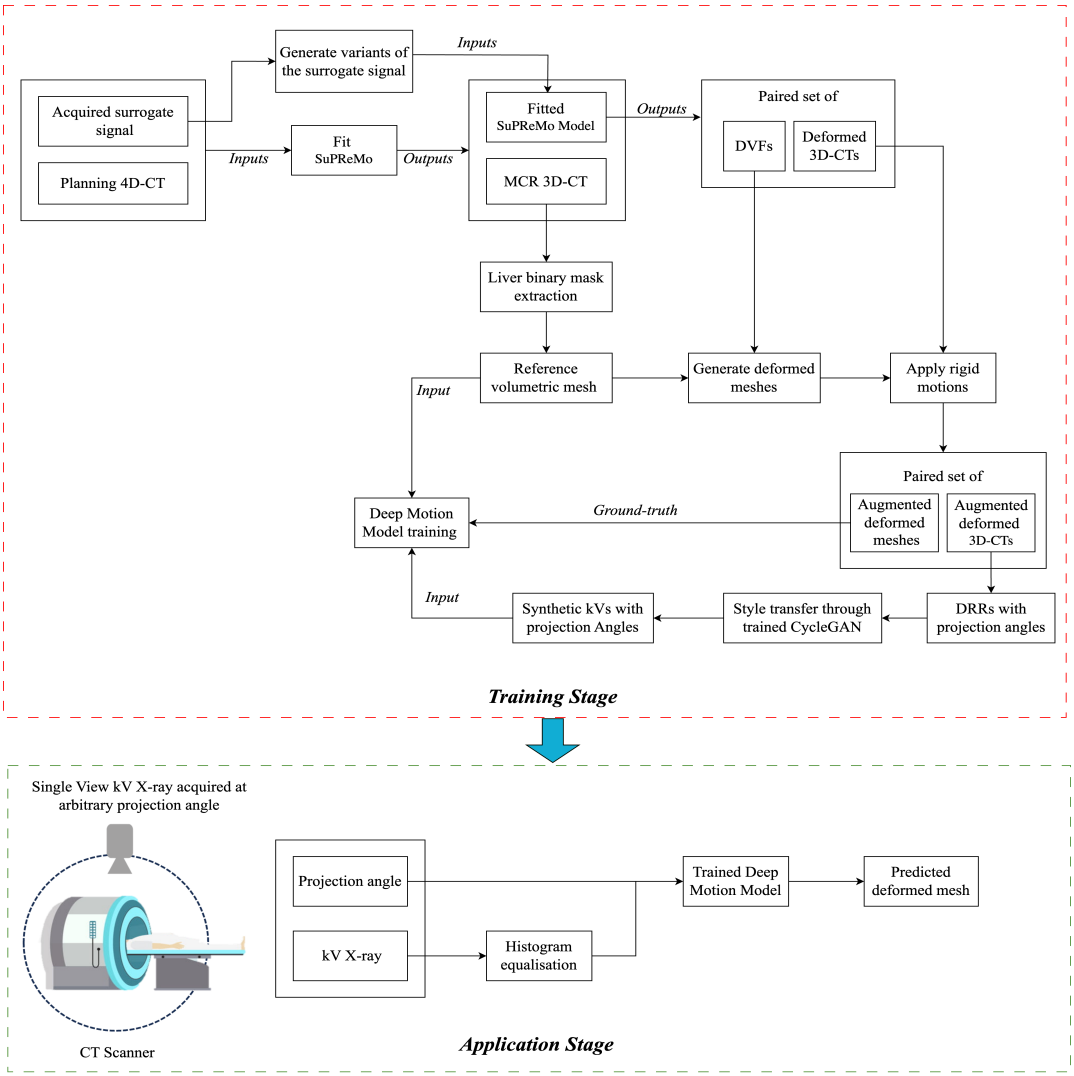


Figure 3.1: During the training stage, a synthetic motion dataset was created and a GNN model was trained to predict volumetric liver mesh deformation from a single X-ray projection. In the application stage, the trained GNN model was employed to derive the predicted deformed mesh using a real kV X-ray image captured at any arbitrary projection angle during the treatment process.

3.2 Graph Neural Networks

This section provides a concise overview of the background pertaining to GNNs in the spatial domain, specifically focusing on graph-based convolution. Typically, this convolution is depicted as a neighbourhood aggregation or message-passing scheme, aiming to extend the convolution operator's applicability to irregular domains. The concept of message passing within a graph proves to be a robust and influential idea, offering insights into numerous graph algorithms. In essence, it involves nodes within a graph sending and receiving messages through their connections with neighbours. This process can be conceptualized in two steps: firstly, nodes transmit a message describing themselves to neighbouring nodes, and subsequently, the receiving nodes gather these messages to update themselves, gaining a better understanding of their environment. Conceptually, the iterative propagation of input features at each graph node begins with an initial step. This step is then updated by incorporating messages received from connected nodes in the preceding iteration, forming a repetitive cycle where each node starts with its updated state, considering information from its neighbours. The message passing GNN can be represented as:

$$\mathbf{x}_i^{(k)} = \gamma^{(k)}\left(\mathbf{x}_i^{(k-1)}, \xi_{j\in\mathcal{S}(i)}\,\phi^{(k)}\left(\mathbf{x}_i^{(k-1)}, \mathbf{x}_j^{(k-1)}\right)\right) \tag{3.1}$$

In this equation, the notation k represents the layer index in a GNN, indicating the specific layer at which node i's feature vector is being updated. Each layer k involves updating node features based on their previous layer feature vectors and aggregated messages from neighbouring nodes.

- $\mathbf{x}_i^{(k)}$ represents the updated feature vector of node i at layer k in a GNN.

- $\gamma^{(k)}$ is a differentiable function (such as an MLP or a non-linear activation) that combines the feature vector of the previous layer, i.e. layer $k-1$, of the i^{th} node, represented by $\mathbf{x}_i^{(k-1)}$, with aggregated messages from neighbouring nodes j.
- $\xi_{j \in \mathcal{S}(i)}$ denotes a permutation-invariant function (e.g., sum, max, average) applied to the messages $\phi^{(k)}\left(\mathbf{x}_i^{(k-1)}, \mathbf{x}_j^{(k-1)}\right)$ received from neighbouring nodes j in the node's neighbourhood $\mathcal{S}(i)$.
- $\phi^{(k)}$ is another differentiable function (such as an MLP or a non-linear activation) that computes messages from neighbouring nodes' feature vectors $x_j^{(k-1)}$ at layer $k-1$ to node i at layer k.

For instance, in the Graph Convolutional Network (GCN) [222] layer, the aggregated vector undergoes processing through a densely connected layer (i.e., a fully connected layer). This is another way of expressing the multiplication by a weight matrix, followed by the application of an activation function. Notably, the weight matrix is shared among all nodes in the graph for a given layer. The output from this dense layer serves as the new vector representation of the node. This sequential process is applied to every node in the graph. Each node gathers messages from its neighbours, aggregates these messages, and then passes the resulting vector, along with its current state, through a standard neural network. This yields a new vector or the next state that represents the node. The node is subsequently characterized by this new vector.

In the second layer, the same process is repeated, with the input being the updated vector from the first layer. The row inputs are directed to the first layer, and the output of the first layer serves as the input to the second layer, and so forth. The number of GCN layers imposes an upper limit on how far the signal can travel. For example, with two GCN layers, the message passing operation occurs twice, indicating that the signal

from any particular message can travel a maximum of two hops away from the source node. If a long-range connection is crucial for a given problem, additional GCN layers need to be employed. The size of the node vectors emerging from the GCN layer is determined by the number of units in the dense neural layer.

3.2.1 Graph Attention Networks

In the GCN layer, when features are aggregated, every neighbouring node's features are assigned identical weights concerning the current node of interest. Graph attention networks address this issue differently by incorporating an attention mechanism in each layer. This mechanism assigns different weights based on how the features interact with the current node of interest.

Let $X^{(0)} = \{{x_1}^{(0)}, {x_2}^{(0)}, \ldots, {x_N}^{(0)}\} \in \mathbb{R}^F$ denote the representation of input feature vectors for nodes in a graph G, where N is the number of nodes, and F is the number of features per node. The superscript (0) indicates these features are from layer zero, representing the initial feature vectors of the nodes before processing by any GNN layers. Now, consider passing these initial features through a single GNN layer that computes a new set of node features $X^{(1)} = \{{x_1}^{(1)}, {x_2}^{(1)}, \ldots, {x_N}^{(1)}\} \in \mathbb{R}^{F'}$, where F' represents the length of the output feature vectors per node after the transformation. The superscript (1) denotes the node features at layer 1, indicating the node features after passing through a single GNN layer.

The initial stage of any GNN layer, including the Graph Attention (GAT) layer, involves message transformation. In this phase, a message is produced from each node i by applying a learnable linear transformation function parameterized by a weight matrix $W \in \mathbb{R}^{F' \times F}$ to the corresponding input feature vector ${x_i}^{(0)}$. This linear transfor-

mation is shared across all nodes within the layer. The subsequent step is to compute attention coefficients using a shared attention mechanism A, determining the relative importance of neighbouring characteristics to the current node of interest [223].

$$e_i j = A(W x_i^{(0)}, W x_j^{(0)}) \tag{3.2}$$

This indicates the importance of the features of node j to node i (i.e., the current node of interest) by computing the pairwise unnormalized attention score between node i and j. Initially, the authors [224] concatenate the linearly transformed embeddings of the two nodes i and j. They then pass this concatenated vector through the attention mechanism A, a single-layer feed-forward neural network parameterized by a learnable weight vector $W_A \in \mathbb{R}^{2F'}$, along with LeakyReLU as the non-linear activation. Due to the diverse graph structures, nodes may have a different number of neighbours. Thus, the attention coefficients undergo normalization using softmax activation, as illustrated in equation 3.3, ensuring a consistent scale across all neighbourhoods.

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{q \in \mathcal{S}(i)} \exp(e_{iq})} \tag{3.3}$$

Where $\mathcal{S}(i)$ is the neighbourhood of node i. Subsequently, perform the neighbourhood aggregation step by computing the linear combination of the features corresponding to the normalised attention coefficients and employ these outputs as the network's final features.

$$x_i^{(1)} = \gamma^{(1)}(\sum_{j \in \mathcal{S}(i)} \alpha_{ij} W x_j^{(0)}) \tag{3.4}$$

Here $\gamma^{(1)}$ denotes a non-linear activation function.

Multi-head attention can be used to stabilise the attention process. This allows for the use of different independent attention mechanisms to perform output feature transformation. The outputs from these multiple attention heads are then combined through operations such as averaging or concatenation.

3.3 Group Normalization

The group normalization layer [225] performs normalization on a mini-batch of data across grouped subsets of channels independently for each observation during model training. This technique involves dividing the channels within a layer into groups and computing the mean and standard deviation along the spatial dimensions and across the grouped channels for each observation independently. After calculating the statistics for each group, the activations within each group are normalized using these group-specific statistics. This process ensures that the activations within each group exhibit consistent scale and distribution. Subsequently, the activations within each group are adjusted using learnable scale and offset parameters, similar to batch normalization [226]. These parameters provide the model with the ability to modify the normalized activations during training.

Compared to batch normalization, which computes statistics across spatial and batch dimensions within mini-batches, group normalization exhibits more stable behaviour across different batch sizes. Batch normalization can suffer from inaccurate batch statistics with smaller batch sizes, potentially leading to higher reported errors [225, 208]. In contrast, group normalization's calculation is independent of batch size, ensur-

ing consistent performance and more reliable results regardless of batch size variations. This stability in group normalization contributes to improved model performance and error reduction.

3.4 Spectral normalization

Spectral normalization [227] is a technique used in deep learning, primarily to stabilize the training of neural networks, particularly in the context of GANs. This technique involves normalizing the spectral norm of weight matrices during training. The spectral norm of a weight matrix is defined as the largest singular value of the matrix, which represents its maximum stretch factor. This singular value can be obtained, for example, through singular value decomposition.

During spectral normalization, each weight matrix in the neural network is divided by its spectral norm. For example, for a weight matrix W, spectral normalization scales W by a factor such that $\sigma_{max}(W)$, where $\sigma_{max}(W)$ is the largest singular value of W. This normalization technique helps control the Lipschitz constant (a measure of how much a transformation can stretch space) of the associated transformations performed by the neural network, leading to more stable training and improved convergence, especially in scenarios prone to issues like mode collapse or unstable gradients (exploding and vanishing gradients) in GANs [227].

3.5 Pixel shuffle layer

Pixel Shuffle is an image upscaling technique employed in deep learning methodologies to boost image resolution. This process involves converting a lower-resolution

image into a higher-resolution one using sub-pixel convolutional layers. Each sub-pixel convolution layer is a combination of a convolution and a pixel shuffle operation. These specialized layers are trained to utilize an array of filters aimed at enhancing the resolution of the lower-resolution feature maps.

The Pixel Shuffle layer accepts an input tensor of shape $(N, C \times r^2, H, W)$ where N represents the batch size, C denotes the number of channels, r signifies the upscaling factor (e.g., $r = 2$ for doubling the resolution), and H and W are the height and width of the input feature maps, respectively. The primary objective of the Pixel Shuffle layer is to rearrange the elements within each channel of the input tensor to generate an output tensor of shape $(N, C, H \times r, W \times r)$. This process involves reshaping each $r \times r$ block of elements from the input feature map into a single pixel in the output feature map. Through this reordering and aggregation of blocks, the Pixel Shuffle layer effectively enhances the spatial resolution of the feature map.

3.6 Synthetic dataset generation

Model training requires paired sets of organ motion instances and corresponding kV images. However, to the best of our knowledge, there are no means of directly measuring such motions while also acquiring the requisite images. Hence, we use synthetically generated data to train and, partially, evaluate the model. Plausible patient-specific motion patterns are extracted from 4D-CT images, and new synthetic instances are produced by interpolating and, within reasonable bounds, extrapolating from these. The process is as follows: 1) 4D-CT images are analysed using the SuPReMo toolkit (Surrogate Parameterized Respiratory Motion Model) [199, 228], which produces, inter alia, a model of the motion present in the images, linked with appropriate surrogate

signals; 2) new, yet plausible motion instances are generated from this model by randomly perturbing the surrogate signal; 3) the resulting motion fields are used to deform the reference CT volume; 4) DRRs are generated from these deformed volumes for all required projection angles, and 5) the DRRs are style transferred to match kV image intensity and noise distributions. The result is a set of realistic 'kV' images of the deformed anatomy acquired at various projection angles, for which ground-truth 3D motion states are known. Finally, the target organ is segmented from the reference CT volume and an organ template mesh is constructed. Full details are presented in the following sections.

3.6.1 Overview of the clinical dataset

The Deep-Motion models were evaluated using data from four liver cancer patients, with a focus on liver motion. Each patient dataset comprised: 1) a 4D-CT with 10 phases (i.e. $10 \times$ 3D-CT volumes), spatial resolutions of $0.98 \times 0.98 \times 2.0$ mm^3, and image dimensions of $512 \times 512 \times 105$; and 2) two kV scan series, each covering approx. 4-5 mins of free breathing and a full rotation of the treatment gantry, acquired at the start of treatment sessions on different days. As per clinical practice, the kV scans were centred on the liver region and used a constrained field-of-view (FOV). As a result, for some projection angles, the images include only a segment of the liver.

3.6.2 Generation of synthetic motion states from 4D-CT data

SuPReMo is a toolkit for simultaneously estimating a motion model and constructing motion-compensated images from a 4D-CT dataset. The resulting model describes the anatomical motion present in the raw data over a single (averaged) breath cycle and is

linked with corresponding scalar surrogate signals derived, for example, from breathing traces or image features. For formulation reasons (see [199, 228]), two surrogate signals are required; however, these needn't be independent, and as a practical measure in each case we constructed the second signals as temporal derivatives of the first, computed for example using finite differences. Example surrogate signals are plotted in Fig. 3.2. Our 4D-CT datasets each decompose the breath cycle into 10 bins (phases). With the usage of two surrogate signals, the SuPReMo model possesses two degrees of freedom, enabling it to simulate variable motion, including intra- and inter-breath variability.

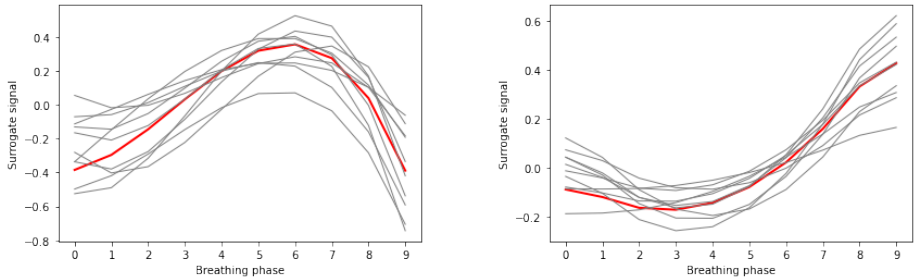


Figure 3.2: Example surrogate signals: original signals associated with the input 4D-CT data (red) and randomly generated variations from these (grey), used in turn to synthesise new motion states. The first and second signals are plotted on the left and right, respectively.

First, we fit SuPreMo's motion model with input surrogate signal and 4D-CT data for each patient case. This returns the motion-compensated reconstructed 3D-CT (MCR) image volume and the fitted motion model, which are then utilised to simulate new motion states by varying the input surrogate signal. Each point s_i, $i \in [0, 9]$ on the curve is randomly perturbed by a value in the range $\pm 0.4 s_i$. Extrapolating beyond this range is more likely to cause unrealistic motion and even folding in the resulting images. A new surrogate signal is then created by fitting a 3$^{\text{rd}}$ order polynomial to

the new points. The latter ensures the new signal remains smooth over the full breath cycle. Examples of such generated signals are shown in Fig. 3.2. Using this new signal, and MCR as a reference volume, SuPReMo's motion model generates corresponding deformed 3D-CT volumes and their related DVFs.

For each test patient case, we created 11 separate surrogate signals, each comprising ten deformed configurations, resulting in 110 synthetic deformation states in total. To introduce more diversity into the synthetic motion instances and substantially deviate from the original 4D-CT data, we incorporated rigid motions by applying random translations/shifts along the LR, AP, and SI directions and produced a total of 550 deformed states. This is also advantageous during testing on real in-treatment kV images, as it reflects the variations in onboard patient setup across different scan series or fractions, which can potentially lead to shifts in the in-treatment kV images.

3.6.3 Generation of synthetic kV X-ray images

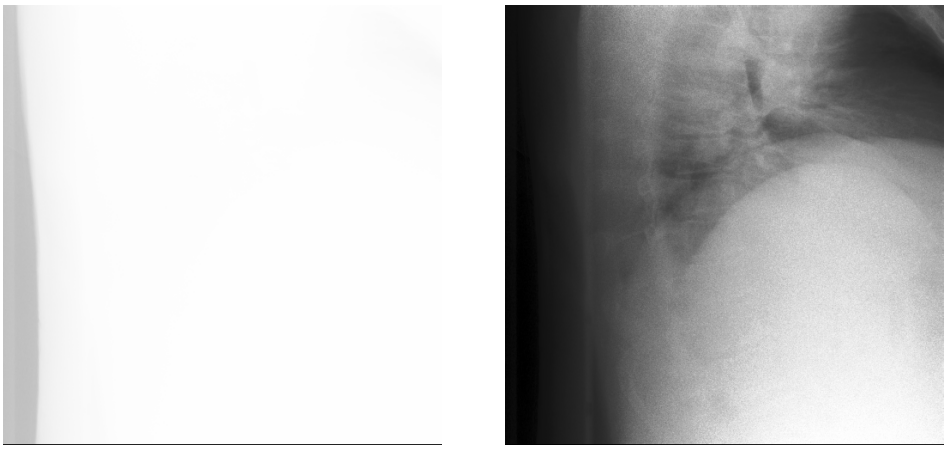
For each deformation state, DRRs were obtained from the deformed 3D-CT volumes using an enhanced version of the Siddon-Jacobs ray tracing technique [185, 186] presented in RTK toolkit [229]. The latter, given a source position and projection direction, computes the line integral of Hounsfield unit densities along ray lines to a prescribed 2D plane. In radiotherapy, in-treatment kV images are acquired perpendicularly to the anatomical axial direction.

The source-to-isocenter (SID) distance was set to 1000 mm, while the SDD was set to 1536 mm. The origin, pixel spacing, and image dimensions in the detector plane were defined based on the Elekta projection configuration, with values of [-204.4, -204.4, 0] in mm, [0.8, 0.8, 0.8] in mm, and [512, 512, 1], respectively.

The following procedure was implemented to mitigate anatomical positioning differences caused by FOV incompatibility during DRR generation. In each patient case, the phase zero 3D-CT from the planning 4D-CT data, chosen from the available ten phases, was aligned with a 3D-CBCT volume reconstructed from one of the kV projection scan series. This alignment utilized rigid registration (translation only). Following this, the same transformation was applied to the other phases in the 4D-CT, ensuring their alignment with each other and the CBCT volume. The SlicerANTs Registration package in the 3D-Slicer toolkit facilitated this alignment process. The SuPReMo model was then fitted using this transformed 4D-CT data before generating the synthetic dataset. Finally, to ensure the proper alignment of the FOV between synthetic and real kV images for each patient, synthetic images were created using the FOV origin header details extracted from the real images in the same kV projection scan series used for aligning the 4D-CT data.

However, the resulting DRRs lack scatter properties and noise characteristics of genuine kV X-ray images, leading to a sharper and higher contrast appearance. This discrepancy arises from the fundamental ray-tracing method used in DRR algorithms, which differs from the X-ray beam utilized in conventional CT scanners. While DRR algorithms model only the attenuation of primary photons, the generation of real kV projection images is influenced by additional physical phenomena such as beam-hardening [230].

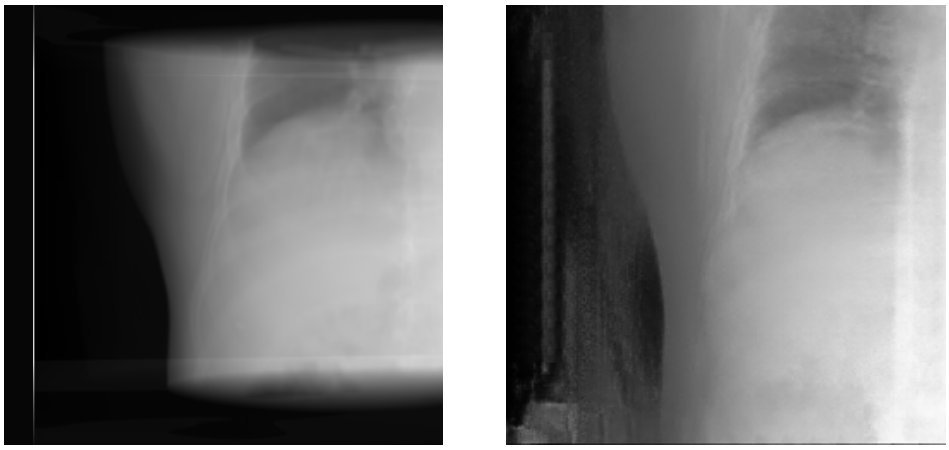
Hence, it is necessary to perform a style transfer from real kV X-rays to DRRs. However, it is impossible to acquire paired DRR and real kV X-ray images from the same patient or scenario. Therefore, the only viable solution is to utilize an unpaired image-to-image style translation technique to transfer styles from real kV to DRRs. These



(a) Before histogram equalization (b) After histogram equalization

Figure 3.3: Visualization of a real kV image before and after applying histogram equalization

techniques enable us to utilize existing datasets of DRRs and real kV X-ray images without the need for explicitly matched pairs. Consequently, a Cycle-GAN model [231], conditioned on the projection angle, was trained on an unpaired set of real kV X-ray images and DRR images. This model allows the transfer of style from real kV X-rays to the DRRs, enabling them to mimic synthetic kV X-rays. This process is conducted individually for each case since FOV acquisition varies from patient to patient. The image contrast in the kV projections was enhanced using the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique [232], with a clipping level of 5 and a tile grid size of 3×3 (see Figure 3.3). By preprocessing low-contrast images with histogram equalization improves model training outcomes by facilitating more effective learning of relevant patterns and structures, potentially leading to enhanced accuracy and performance [233]. Subsequently, all DRRs were passed through the trained conditional CycleGAN to generate the final synthetic kV X-ray images (see an example in Figure 3.4).



(a) A DRR image (b) Corresponding synthetic kV

Figure 3.4: Visualization of a DRR image and its corresponding synthetic kV image

Conditional CycleGAN architecture

The architecture consists of two GANs to map between two image distributions. Each GAN has its own generator and a discriminator. The generator of the first GAN (G_{DRR2kV}) generates synthetic kV from a given DRR and its discriminator (D_{kV}) distinguishes the synthetic kV from the real kV. The generator of the second GAN (G_{kV2DRR}) generates DRR from a given kV whereas the discriminator (D_{DRR}) distinguishes the synthetic DRR from real DRR. Each GAN in our architecture is a conditional GAN, the condition of which is based on the projection angle of the input image for both the generator and the discriminator. As in the original CycleGAN paper [231], the total loss to train our conditional CycleGAN is composed of three components: adversarial loss, cycle consistency loss, and identity loss as described below.

Adversarial loss This loss function serves as a pivotal mechanism for training each generator to produce data that closely resembles real data, while simultaneously refining the corresponding discriminator to more effectively distinguish between real and

generated data. During the training process, a dynamic interplay unfolds between the generator and discriminator, with each being alternately trained while the other remains fixed. A detailed breakdown of this interactive process is as follows:

1. When the generator successfully misleads the discriminator (i.e., the discriminator fails to classify a generated image as fake), it signals a need for the discriminator to enhance its discriminatory capabilities. Consequently, the loss is back-propagated through the discriminator to facilitate this enhancement.

2. Conversely, if the discriminator accurately distinguishes between fake and real images, it prompts the generator to improve its performance. Consequently, the loss is back-propagated through the generator network to encourage this refinement.

The primary goal is for the generator to minimize this adversarial loss against its corresponding discriminator, which undertakings to maximize it. Ultimately, this training regimen aims to enable generators to produce translated images that are virtually indistinguishable from real images within the target domain. By leveraging this loss function, the generator learns to approximate the distribution of training data from the target domain and subsequently samples from this learned distribution.

Cycle constancy loss This loss function measures the discrepancy between the initial input image and the image reconstructed after passing through both generators. For instance, if we translate a DRR to kV by feeding an input DRR image which is represented as $input_{DRR}$ through the G_{DRR2kV} generator, and then translate it back from kV to DRR by passing it through the G_{kV2DRR} generator, ideally, we should return to the original image, $input_{DRR}$. This loss is applied in two ways:

1. Forward Cycle Consistency Loss:

   $input_{DRR} \rightarrow G_{DRR2kV}(input_{DRR}) \rightarrow G_{kV2DRR}(G_{DRR2kV}(input_{DRR})) \approx input_{DRR}$

2. Backward Cycle Consistency Loss:

   $input_{kV} \rightarrow G_{kV2DRR}(input_{kV}) \rightarrow G_{DRR2kV}(G_{kV2DRR}(input_{kV})) \approx input_{kV}$

These losses encourage the generators to produce outputs that are not only realistic but also preserve the content of the original input image after undergoing translation across domains, thus ensuring consistency.

Identity loss This loss function serves a critical role in guiding the generator's output to closely match the original input when the input belongs to the target domain. This loss function works by minimizing the difference between the generated output and the input image, thereby preventing excessive distortion of images during the transformation process.

Conditioning with projection angle Both the generator and discriminator in each GAN take an image and its corresponding projection angle as inputs. The projection angle was normalized by dividing it by 360 since the gantry rotation varies from 0 to 360 degrees. Since the projection angle is represented by a scalar value, the first step involves creating a 2D matrix that matches the dimensions of the input image. In this matrix, each element is assigned the scalar value corresponding to the projection angle. Subsequently, the resulting matrix is concatenated with the input image as an additional channel before undergoing forward propagation through either the generator or discriminator.

Generator configuration Both generators have the same architecture and take, as input, images of size 256×256. The encoder part or the downsampling of the generator consists of four convolutional layers with 64, 128, 256, and 512 filters, respectively. The kernel size was set to 4×4, stride to 2, and padding to 1 for all convolutional layers. We then added two more convolution layers with 1024 and 512 filters without any downsampling operation by initializing them with a kernel size of 3×3, stride, and padding both set to 1. The decoder part of the generator comprises four upsampling layers where each layer has a convolutional layer followed by a pixel shuffle layer which is used to upsample the input image [234] since we encountered checkerboard-like artifacts in several cases. The convolution layers in decoder have 256×2^2, 128×2^2, 64×2^2 and 1×2^2 filters, respectively. The kernel size was set to 1×1, stride to 1 without any padding for all convolutional layers. The corresponding pixel shuffle layer rearranges the output feature map with dimensions $H \times W \times (C \times 2^2)$ to a tensor of shape $(H \times 2) \times (W \times 2) \times C$, where 2 is an upscale factor. Rectified Linear Units (ReLUs) were applied for non-linearity after every convolutional operation. Hidden layer outputs were normalized using spectral normalization [227] since this helps to overcome the instability due to vanishing and exploding gradients during the model training.

Discriminator configuration The discriminators consist of five convolutional layers with 64, 128, 256, 512, and 1 filter, respectively. The kernel size was set to 4×4, stride to 2, and padding to 1 for the first four convolutional layers. For all of these four layers, we used LeakyReLU with a negative slope of 0.2 as the non-linear activation. For the last convolutional layer, the kernel size was set to 4×4, stride to 1 with no padding. Hidden layer outputs were normalized using spectral normalization [227].

Discriminators take a 256×256 image as input and produce a 30×30 tensor as output. The classification result for a 70×70 area of the input image is stored in each element of the output tensor. The discriminator checks whether every 70×70 area (these areas overlap each other) of the input image appears real or fake by returning a tensor of size 30×30. The overall classification result is then calculated by taking the average on the 30×30 values.

Training details The model underwent training using an unpaired dataset, which included 11,000 images of DRRs and 6,700 images of real kV images. Correspondingly, the test sets comprised 2,000 DRR images and 1,000 real kV images. The adversarial loss was computed using hinge loss, while both identity loss and cycle consistency loss employed L1 loss. To maintain a balanced weighting among the various loss terms within the total loss function, the identity loss was multiplied by a factor of 5, whereas the cycle consistency loss was multiplied by a factor of 10. A Spectral Normalization layer was applied to all layers in both the Discriminator and Generator. The learning rate for generators and discriminators was set at 0.0002, and the batch size was fixed at 1. The learning rate was maintained at the same value for the initial 50 epochs and subsequently linearly decayed to zero over the following 50 epochs.

3.6.4 Creation of template meshes

Binary masks of the relevant organs were extracted from reference 3D-CT volumes using 3D Slicer's Segment Editor. These masks were then used to generate tetrahedral meshes using the Iso2Mesh [235] tool in MATLAB. However, Iso2mesh focuses exclusively on the voxel coordinates of the image volume and generates mesh coordinates in voxel units, assuming isotropic voxels with dimensions of $1mm \times 1mm \times 1mm$

(unit voxel spacing). Additionally, Iso2mesh disregards the image offset (i.e. image origin) and derives node positions relative to the $(0, 0, 0)$ point in physical coordinates. Therefore, we first rescaled the resulting mesh coordinates by the true voxel spacing ($i.e., 0.98mm \times 0.98mm \times 2mm$) and then translated the mesh coordinates to accommodate the offset present in the CT image volume. This means the meshes are physically aligned with the relevant anatomical regions in the 3D-CT volumes. The extracted binary mask (in green colour) and corresponding volumetric mesh overlaid on the axial, coronal, and sagittal planes within the reference 3D-CT image volume are shown in Figure 3.5.

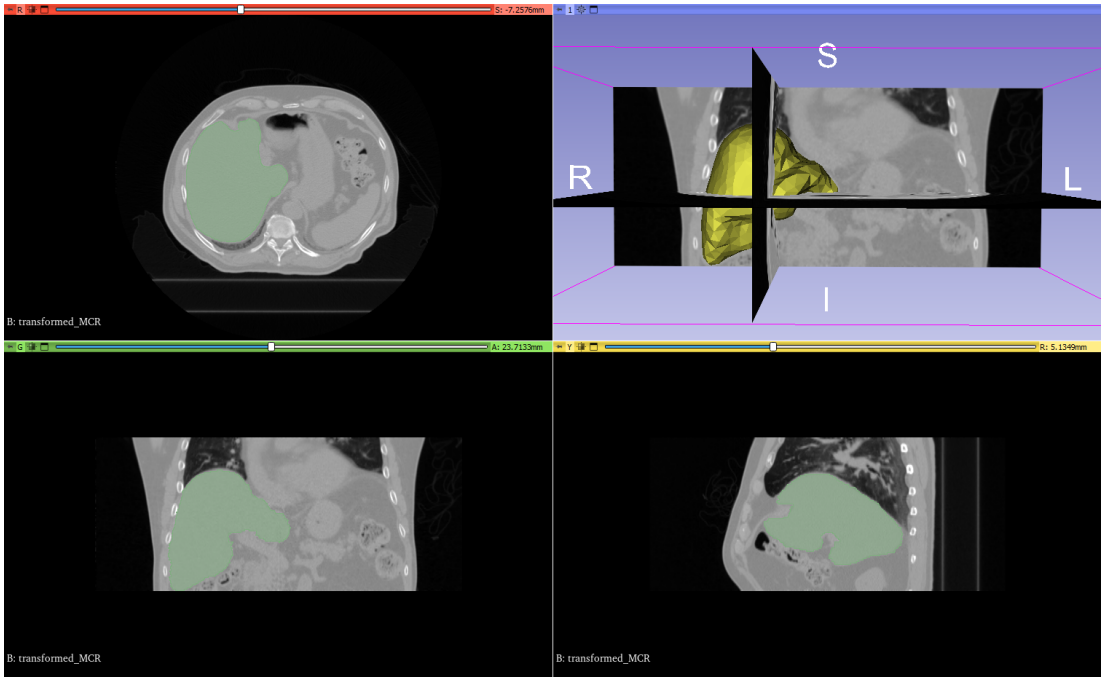


Figure 3.5: Liver binary mask extraction from a reference 3D-CT and the corresponding volumetric mesh alignment on axial, coronal and sagittal planes

3.6.5 Deformed Volumetric Meshes Generation

Ground-truth positions of the mesh nodes for the generated deformation states, which are the deformed 3D-CT image volumes, are obtained by interpolating the displacement vector field at the node positions. The associated displacement vector field was extracted from the corresponding DVF. The transformation defined in each displacement field was then applied to every corresponding mesh point in the physical space.

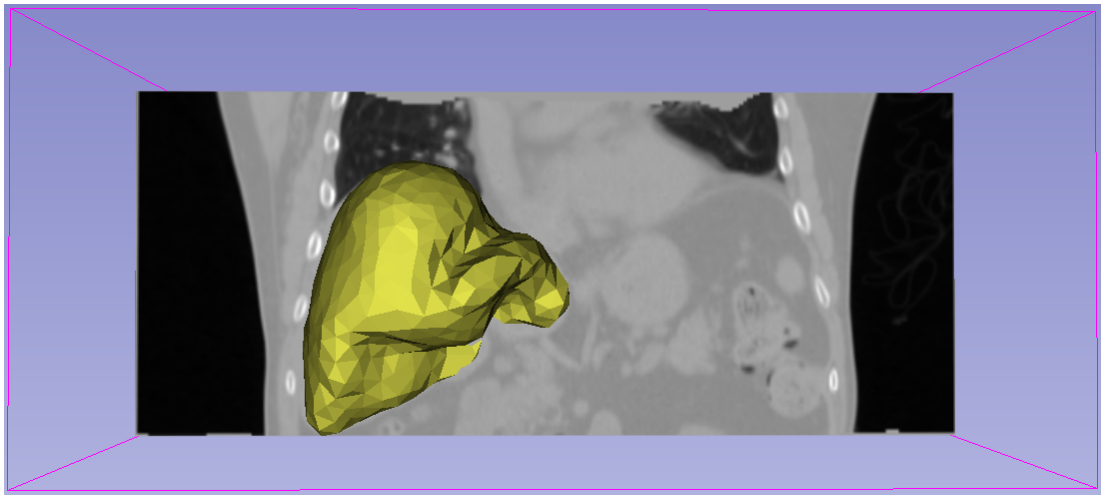


Figure 3.6: A deformed liver mesh alignment with corresponding deformed 3D-CT volume

3.7 Methodology

Graphs serve as a natural representation for organ geometries due to their adeptness at effectively capturing shape and topology, including local connectivity variations [236]. Recent advancements in utilizing GNNs for 3D shape reconstruction from single-view images, as demonstrated in studies such as [215, 217, 218, 220, 221], highlights the state-of-the-art capabilities of GNNs in processing complex 3D organ geometries. GNNs are proficient in preserving critical topological features such as node connec-

tivity, which enables to define higher-order loss functions like Laplacian loss. These loss functions are useful in regularizing 3D shapes, ensuring geometric fidelity that is otherwise challenging to achieve without the explicit graph representation provided by GNNs [237]. The inherent suitability of GNNs for encoding graph-based representations of organ geometries forms a foundational rationale for their integration into our approach. By this means the core of our approach is a GNN that learns mappings from kV image features to nodal displacements of a patient-specific organ mesh. The model is trained individually for each patient. This section provides a comprehensive description of the components of the approach, including the 3D organ representation, model architecture, and loss functions.

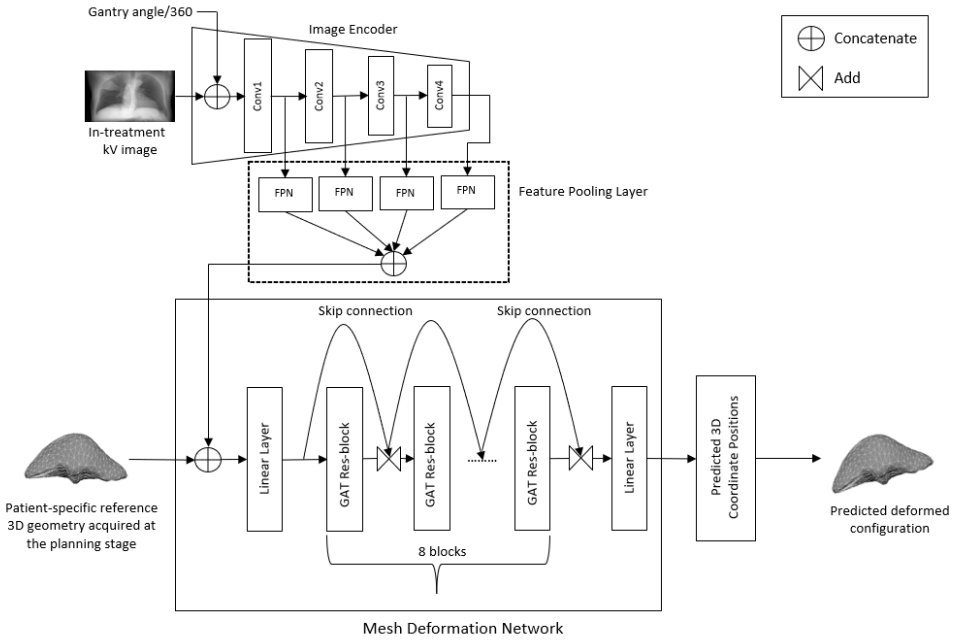


Figure 3.7: Illustration of the Deep-Motion-Net architecture. A 2D-CNN image encoder extracts projection angle-dependent semantic features from an input kV X-ray image. A feature pooling layer comprising four learnable feature pooling networks attaches these features to the appropriate vertices in the patient-specific template mesh. Finally, a graph-attention-based network predicts the corresponding mesh deformation.

3.7.1 3D organ shape representation

We use a 3D unstructured tetrahedral mesh to describe the volumetric organ shape, rather than its surface only. The mesh can be described as an undirected graph $G = \{V, E, F\}$, where V is the set of N vertices in the mesh, E represents the set of edges between connected vertices, and F are feature vectors attached to vertices. Patient-specific organ template meshes, derived from reference CT volumes, are constructed for each patient. In our experiments, we used meshes with N = 785, 827, 803, and 756, respectively, for livers in four patients.

3.7.2 Model architecture

The proposed architecture (shown in Figure 3.7) consists of two components: a 2D-CNN encoder with four learnable feature pooling networks (FPNs), which extract perceptual features from the kV image and attach them to mesh nodes and a GAT-based mesh deformation network.

Incorporating projection angle

Inputs to the 2D-CNN encoder are a single-view kV image and its corresponding projection angle. We first normalise the projection angle by dividing it by 360 since the gantry rotation varies from 0 to 360 degrees. Since the projection angle is represented by a scalar value, the first step involves creating a 2D matrix that matches the dimensions of the input image. In this matrix, each element is assigned the scalar value corresponding to the projection angle. Subsequently, the resulting matrix is concatenated with the input image as an additional channel before it is sent to the image encoder.

2D-CNN configuration

The projection angle-dependant perceptual features of the image are then extracted using four convolutional layers in the image encoder, containing 16, 32, 64, and 128 filters, respectively. For all convolutional layers, the kernel size was set to 3×3, and stride and padding were both set to 1. Exponential Linear Units (ELUs) were applied for non-linearity after every convolutional operation. Hidden layer outputs were normalised using group normalisation [225] since this helps reduce the internal covariate shift, which regularly alters the distribution of the hidden-layer activations during model training. Output feature maps of each convolutional layer were then down-sampled using 2×2 max-pooling layers with stride two before passing into the next layer.

Feature pooling networks

One important challenge encountered during the projection of the template organ geometry onto the motion-compensated kV image is the potential misalignment of projected points, leading to inaccurate feature extraction from the intended anatomical locations due to organ displacement. This discrepancy affects the precision of feature extraction, where expected anatomical features may not align properly with the projected points on the image. Additionally, in clinical scenarios, kV X-ray projections often have a limited FOV, which poses challenges for feature extraction when projecting vertices onto the DRR plane. This is due to the potential for projected mesh nodes to extend beyond the FOV of the projection, leading to difficulties in extracting features accurately. These limitations are challenging as they hinder the CNN encoder from learning optimal features that would enhance the predictive performance of the down-

stream graph network. To overcome these limitations, we introduced learnable feature pooling networks (FPNs) which effectively learn the optimal association of features with mesh nodes. This solution enables end-to-end training of the network, thereby enhancing the overall model performance. We use four FPNs, each coupled with its respective CNN convolutional layer (Figure 3.7). Each FPN has two layers: an adaptive average pooling layer (AAP) and a FC layer. The AAP, with output size 7×7, is applied over the output of the corresponding convolutional layer to reduce the dimension so that an input feature map with dimensions $Height(H) \times Width(W) \times Channels(C)$ is reduced to $7 \times 7 \times C$. The output is flattened before sending it into the FC layer, which contains $5N$ neurons. We reshaped the output feature vector of each FPN into (N, 5). These outputs were then concatenated with 3D coordinates of the template mesh before feeding into the graph network. Each mesh vertex thereby acquires a total of 23 features: five from each of the four FPNs, and three representing the 3D vertex position.

Mesh deformation network

The objective of this network is to estimate the 3D coordinates for each vertex in the deformed mesh configuration. We developed a GNN architecture which relies on a series of GAT-based blocks with residual connections (i.e. skip/shortcut connections). We applied residual connections to increase the impact of earlier layers on the final node embeddings. These connections substantially speed up training and produce better-quality output shapes. The fundamental building block is identical to the Bottleneck residual block [238], with 1×1 convolution and 3×3 convolution layers replaced by per vertex FC layers and GAT layers [223], respectively. Further, group normalization layers are used instead of batch normalization [226] since small batch sizes result in

incorrect estimates of batch statistics, which substantially increases model error. We employed ELUs to impart non-linearity.

This deformation network is similar to the graph-CNN architecture described in [208] but with two differences: graph convolutional network (GCN) layers are replaced by GAT layers and ReLU is replaced by ELU activation. GAT layers incorporate attention mechanisms which allow the assignment of different weights for different neighbouring node feature vectors depending on how they interact. GCNs [222], by contrast, use isotropic filtering and thereby assign similar weight to all feature vectors around the current node. ELU activation avoids the dying ReLU problem and may generate negative non-linear outputs.

3.7.3 Loss functions

The overall objective function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{Shape} + \lambda \mathcal{L}_{Laplacian}, \tag{3.5}$$

with weighting term λ. $\mathcal{L}_{Shape}$ quantifies the difference between the predicted and ground-truth 3D meshes. The template mesh starting shape is defined by its vertices V (Sect. 3.7.1). We define $Y = \{y_1, y_2, \ldots, y_N\} \in \mathbb{R}^3$ and $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\} \in \mathbb{R}^3$ to be the ground-truth and predicted deformed positions in 3D space of these vertices. An intuitive objective is then to minimize the per-vertex L_1 loss between Y and $\hat{Y}$:

$$\mathcal{L}_{Shape} = \sum_{i=1}^{N} \left\| \hat{y}_i - y_i \right\|_1, \tag{3.6}$$

where y_i and $\hat{y}_i$ are the i^{th} vertices in the respective sets.

Using $\mathcal{L}_{Shape}$ alone the vertices were found to move too freely. We introduced a discrete Laplacian [206] loss $\mathcal{L}_{Laplacian}$ as a regularization term to limit this freedom by ensuring vertices do not move too far in relation to their neighbours. This loss penalizes large variations in vertex positions between neighbours, leading to smoother meshes when minimized during optimization. By minimizing the Laplacian loss, vertices adjust their positions to align more closely with neighbouring vertices, resulting in a mesh that exhibits reduced sharp edges or abrupt changes in surface geometry. This regularization effect, known as the shrinkage effect, contributes to the development of smoother and more uniform mesh surfaces. The discrete Laplacian of a vertex with 3D position $\hat{y}_i$ is denoted as:

$$\delta_{\hat{y}_i} = \frac{1}{|\mathcal{S}(\hat{y}_i)|} \sum_{j \in \mathcal{S}(\hat{y}_i)} (\hat{y}_i - \hat{y}_j), \tag{3.7}$$

where $\hat{y}_j$ is a neighbouring vertex of $\hat{y}_i$ and $\mathcal{S}(\hat{y}_i)$ denotes the set of all such neighbours. The discrete Laplacian loss is then given by:

$$\mathcal{L}_{Laplacian} = \frac{1}{N} \sum_{i=1}^{N} \left\| \delta_{v_i} - \delta_{\hat{y}_i} \right\|_2^2, \tag{3.8}$$

where δ_{v_i} and $\delta_{\hat{y}_i}$ are the discrete Laplacian before and after the deformation, respectively.

In our experiments, we used $\lambda = 0.1$ to balance the weights of the two losses.

3.7.4 Implementation and training details

The complete network was implemented using PyTorch and PyTorch-geometric [239]. The model was trained using the Adam optimizer, with a learning rate of 0.0002 and

weight decay of 0.001. The batch size for both training and validation datasets was set to 16. Group normalization was used to normalize the hidden layer outputs of both 2D-CNN and GNN. We used single-headed attention in the GAT layers due to memory restrictions. Early stopping was employed to monitor the validation loss with patience of 30 consecutive epochs, and if the loss did not improve, the optimizer itself stopped the training. However, if there was no progress after eight consecutive epochs, the learning rate was decreased by a factor of 0.8. All weights were initialized using the scheme in [240]. The gradient descent converged after 400 epochs to the optimal solution. As described in Section 3.9, we conducted a series of experiments to determine the optimal network components. The training lasted approximately 36 hours with 256×256 image resolution on a Nvidia Quadro RTX 4000 GPU using a Precision 7820 Tower XCTO Base workstation.

3.8 Model evaluation and results

The performance of the comprehensive framework was assessed through the analysis of liver motion, employing clinical data obtained from four liver cancer patients. The model underwent both quantitative testing on synthetic respiratory motion scenarios and qualitative evaluation using in-treatment kV images acquired throughout a complete scan series for liver cancer patients, as detailed in Chapter 3.6. The experimental setup for synthetic motion instances is outlined in Section 3.8.1, while the assessment of real kV images for each patient case is presented in Section 3.8.2. Additionally, Section 3.8.3 provides a thorough comparison with the recently introduced IGCN model [217, 218], elucidating the distinctions between our approach and theirs. The last section (i.e. Section 3.9) presents the ablation experiments we performed.

3.8.1 Experiments on synthetic data

We first evaluated the framework's ability to recover motion states synthetically generated from the clinical data, and for which ground-truth deformations were correspondingly available. For each patient, synthetic motion and image data were generated as per Section 3.6. Each patient's 550 deformation states were then split into training, validation, and test sets in the proportions 350, 100, and 100, respectively. For training and validation deformation states, 100 uniformly sampled (i.e. at projection angle intervals of 3.6°) synthetic kV images were generated (giving 35,000 training and 10,000 validation images). For test states, 50 kV images were generated at *randomly* sampled projection angles (giving 5,000 test images). To ensure good FOV alignment of the synthetic and real kV images for each patient, each synthetic image was generated using FOV origin header information from a corresponding (i.e. acquired at a comparable projection angle) real image within one of the scan series. To avoid doubt, the real kV images subsequently played no further role in the experiment. Model-predicted and ground-truth 3D liver shapes were compared for all synthetic kV images in the test sets.

Summary results are presented in Table 3.1. The Euclidean distance was used to evaluate the distance errors between ground-truth and estimated shapes. Distributions of mean and peak errors for each test set, and for each projection angle (divided into bins), are shown in Figure 3.8. Finally, surface renderings of the ground-truth, reference, and predicted mesh shapes, overlaid on deformed 3D-CT volumes, are presented in Figure 3.9.

For each of the four test sets the overall mean error was low: ≤ 0.22 mm. Within each test set, the maximum peak errors (i.e. overall max error found in any of the deforma-

Table 3.1: Summary statistics for all test sets with synthetic kV images. Patient case numbers are indicated in column 1. E_{pred} and U_{GT} refer to prediction errors and underlying ground-truth nodal deformation magnitudes, respectively. *Mean (std)*: means (and standard deviations) of values across all nodes, all deformation states, and all projection angles. *Mean peak*: means of the peak values for each deformation state across all projection angles. *Max peak*: overall maximum values from all nodes, deformation states and angles. *99th Percentile*: 99$^{\text{th}}$ percentile values from all nodes, deformation states and angles. All values reported in mm.

Case		**Mean (std)**	**Mean peak**	**Max peak**	**99$^{\text{th}}$ Percentile**
1	E_{Pred}	0.16±0.13	1.39	6.75	0.75
	U_{GT}	10.18±1.33	14.71	28.12	13.85
2	E_{Pred}	0.18±0.19	1.99	7.97	0.97
	U_{GT}	11.65±1.76	15.76	34.91	14.38
3	E_{Pred}	0.22±0.34	3.29	14.66	1.81
	U_{GT}	14.89±2.54	19.36	49.63	17.65
4	E_{Pred}	0.12±0.11	1.16	4.36	0.51
	U_{GT}	10.07±1.13	12.86	25.64	11.97

tion states and at any node) were rather higher, ranging from 4.36 mm for patient 4 to 14.66 mm for patient 3. These peak values occurred for the deformation states with the highest ground-truth displacement (namely: 28.12, 34.91, 49.63, and 25.64 mm for patients 1, 2, 3, and 4, respectively) for each test set. Higher underlying ground-truth displacements corresponded, *in general*, with higher peak errors, although mean errors were consistently low. As indicated by Figure 3.8, the displacement prediction accuracy was almost independent of the image projection angle: box-and-whisker plots of both mean and peak errors are very similar across the range of angles.

Finally, it is important to note that the higher peak errors in all cases were extremely localised within the meshes. This is indicated firstly by the low overall mean values and more so by the 99$^{\text{th}}$ percentile errors (Table 3.1), which were below 1 mm for 3/4 test sets and below 2 mm for the fourth. That is, the errors were low, even sub-millimetres, for the vast majority of the mesh nodes. The point is further illustrated by

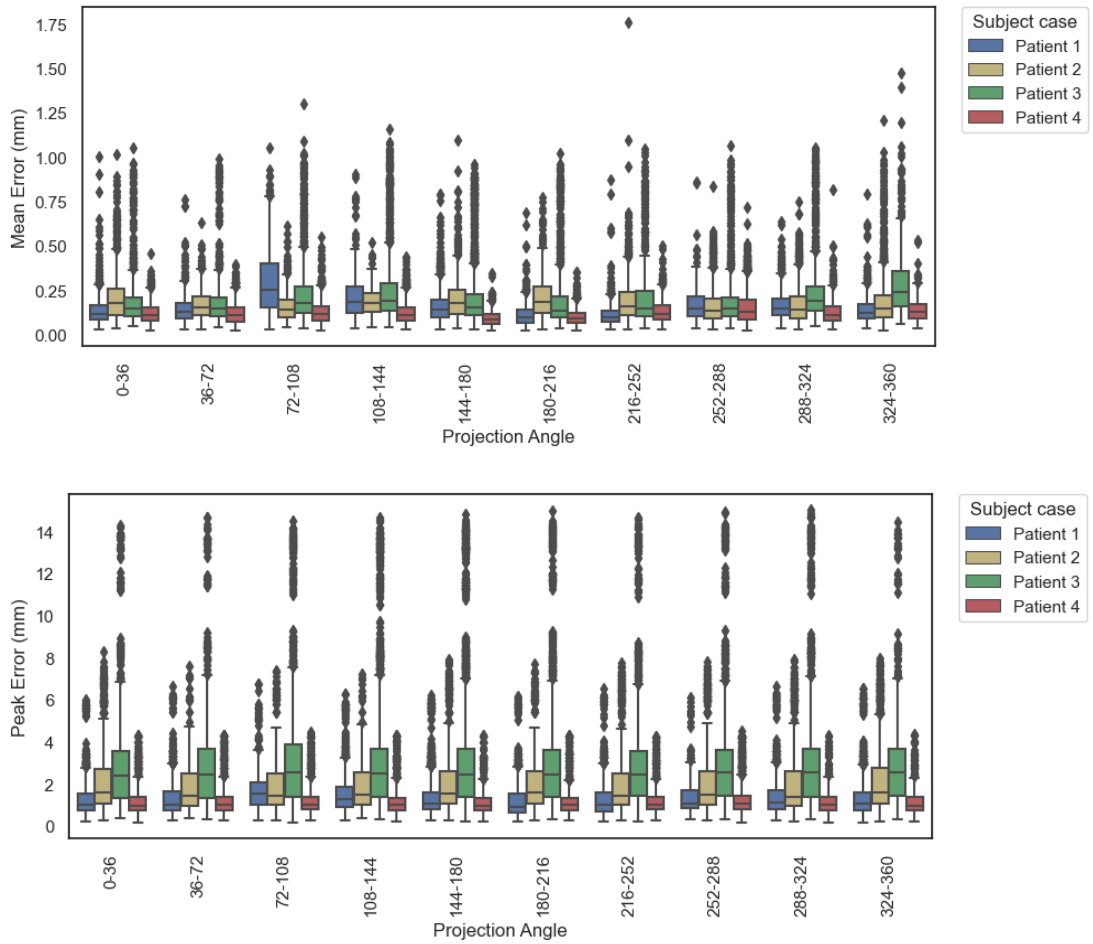
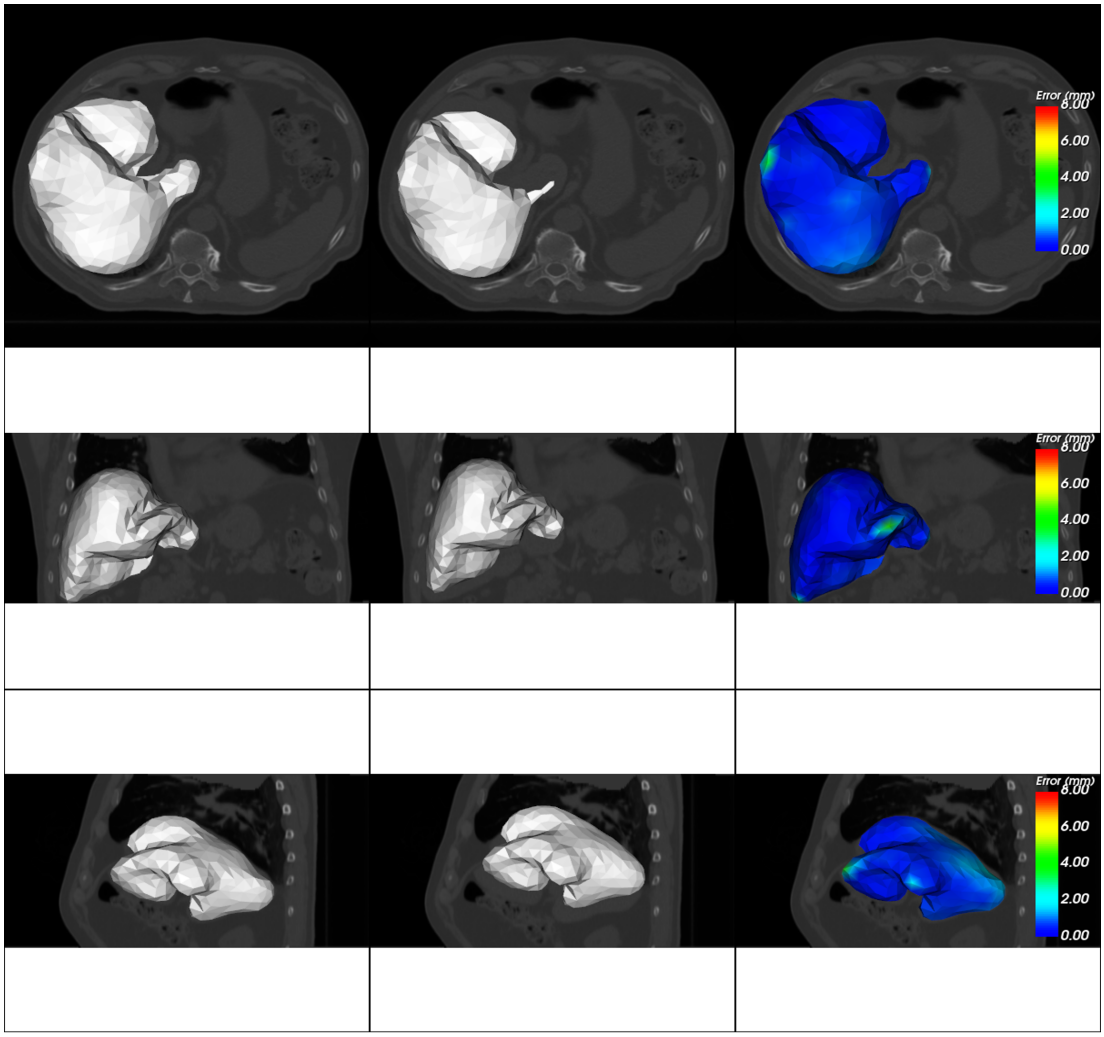


Figure 3.8: Effect of projection angle on prediction accuracy: box and whisker plots of the mean (top) and peak (bottom) prediction errors grouped according to image projection angle (degrees). Each box and whisker shows the distribution of errors for the indicated projection angle using all deformation states in the test set. For clarity of visualisation, angles are further grouped into 10 equal bins covering a full revolution. Results for patients 1 (blue), 2 (yellow), 3 (green), and 4 (red) are shown for each bin.

the renderings of predicted mesh shapes colour-mapped by displacement error in the right columns of Figure 3.9. For liver meshes, the error was close to zero over most of the surface and had only very localised regions of higher values.



GT Mesh Alignment | Reference Mesh Alignment | Estimated Mesh at Projection Angle 80.849

Figure 3.9: Visualisations of ground-truth deformed (left column), template (middle column), and estimated deformed (right column) 3D liver shapes. Meshes are overlaid on the deformed 3D-CT volume. Rows 1-3 show, respectively, axial, coronal, and sagittal views. Results are shown for the *worst* performing test case for patient 1: image projection angle 80.849°, and deformation state producing highest errors. Contours in the right column indicate the spatial distribution of errors on the surface. Similar results for patients, 2-4 are presented in Appendix A.

3.8.2 Evaluation on real kV images

Our second set of experiments used real in-treatment kV images from each patient's second scan series (i.e. the series *not* used during training data creation). These series contained 1378, 1310, 1320, and 1292 images for patients 1, 2, 3, and 4, respectively. In the absence of ground-truth deformations, direct assessment of prediction errors is impossible. Therefore, two approaches were adopted: 1) semi-quantitative assessment based on an image similarity metric between input real kV images and model-generated DRRs; and 2) qualitative assessment based on overlaying model-predicted liver boundaries on input kV images. For the qualitative assessment, all images in the scan series were used. To reduce computation time (associated, in particular, with spline deformation of the image volumes), only 100 images, uniformly sampled, were used from each patient's series in the similarity-based assessment.

Mutual Information-based assessment

3D organ deformations predicted for a given input kV image can be used to deform the patient's reference CT volume. The correspondence between the input image and a DRR generated from this deformed CT volume should then improve with the accuracy of the model prediction. With this in mind, we used kV-to-DRR image similarity, quantified using mutual information (MI), as a surrogate measure of the model's deformation prediction accuracy. In particular, we assessed the improvement in MI when using the model-deformed CT volume compared with using the undeformed volume. The kV and DRR images arise from different modalities and exhibit varying intensity levels for the same underlying structure. MI can capture statistical dependencies and information shared between images, even when they originate from different modal-

ities[241, 242]. It is sensitive to underlying intensity patterns and structural differences across multi-modalities, making it well-suited for evaluating image similarity in scenarios where images exhibit disparate appearances owing to varying acquisition methods or inherent characteristics [243].

The process is summarised in Figure 3.10. The input kV image is passed to the model, which predicts the corresponding 3D organ mesh deformation. A thin-plate-spline (TPS) transformation is initialised using the reference and deformed mesh and used to deform the reference 3D-CT volume. A DRR is then generated using the input kV image's projection angle. Separately, the deformed mesh is used to generate a 3D binary mask, covering the liver-predicted shape, from which a 2D 'mask DRR' is produced. The latter is used to create ROIs in both the DRR described above and the input kV image. MI is then computed between these two ROIs. The ROI masking process ensures the similarity computation is restricted to the image regions to which the model predictions apply; deformations in the remainder of the 3D-CT volume, derived from the TPS transformation, are merely extrapolated rather than directly predicted. Finally, a comparable process is followed to produce a reference DRR (i.e. without accounting for the motion) for comparison: the DRR is generated from the undeformed (reference) 3D-CT volume, an ROI mask is created from the reference organ mesh, and MI is computed between the masked DRR and input kV image.

Table 3.2 summarizes the results of these experiments. The similarity score values using the reference 3D-CTs are lower than for the motion-corrected (deformed) volumes, suggesting our model is making sensible predictions of the liver motion. Moreover, the marginal disparities in average MI values led us to conduct a one-way ANOVA test for each patient case. The resulting p-values for MI differences between the reference

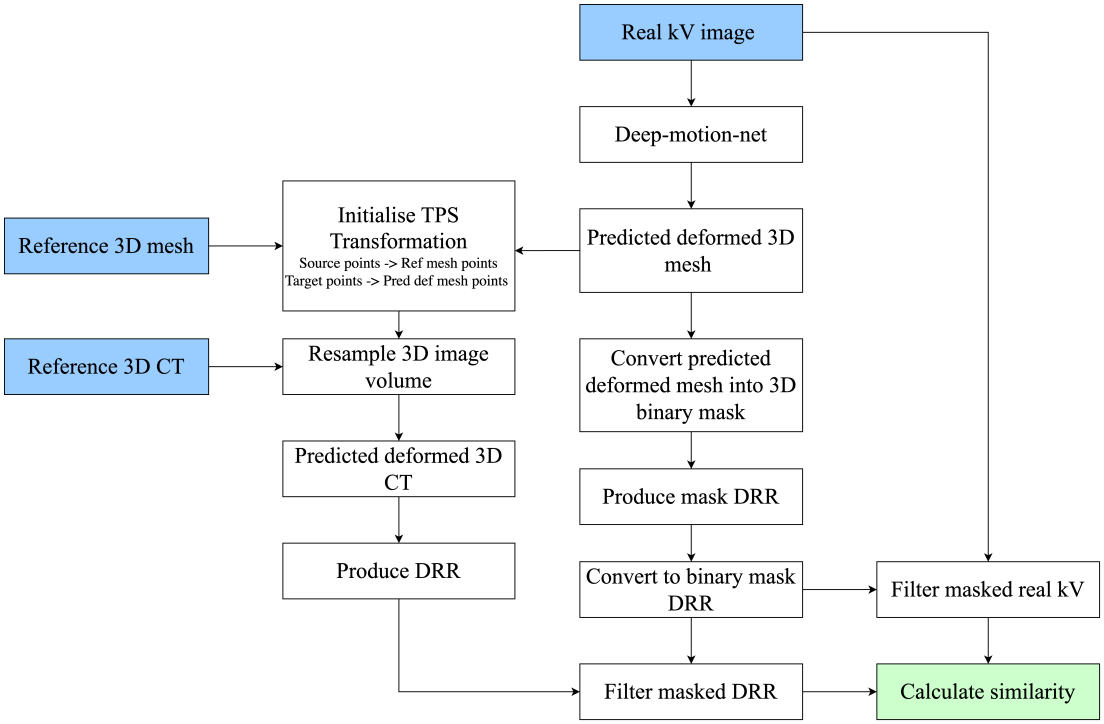


Figure 3.10: Illustration of the process of MI-based assessment of model prediction accuracy.

and deformed versions were 0.0449, 0.0283, 0.0453, and 0.0423 for patients 1, 2, 3, and 4, respectively, suggesting statistically significant (assuming alpha value of 0.05) differences.

Qualitative assessment by boundary overlay

Histogram-equalised kV images from the mentioned scan series were fed into the trained models to obtain 3D mesh shape predictions. As mentioned, all images in the scan series were used. From each predicted mesh, a corresponding binary image volume was generated. Finally, the projected liver surface boundaries were obtained through ray-tracing on these binary image volumes, and superimposed on the respective input kV images. Samples from each patient are shown in Figure 3.11. Supple-

Table 3.2: MI similarity scores (mean ± standard deviation, computed from the 100 images sampled from each scan series) between real kV images and DRRs generated at the same projection angles for each patient case. Column 2 presents values when the reference (i.e. undeformed) CT volume is used. Column 3 presents values when the CT volume is deformed using the model-predicted deformation fields. All values were computed on the liver region.

Case	Reference	Deformed
1	1.16±0.31	1.33±0.23
2	1.14±0.21	1.39±0.14
3	1.13±0.27	1.28±0.23
4	1.31±0.25	1.47±0.15

mentary materials, moreover, include animated versions based on the full scan series of each patient case for further reference. These latter most effectively show (qualitatively) the consistent alignment of the model predictions with the input images across many breathing cycles and the full rotation of the treatment gantry.

3.8.3 Comparison model

We compared the performance of our approach with that of the recently presented IGCN model [217, 218]. As described, like our approach, IGCN predicts 3D organ shapes from single-view X-ray images. It is, however, limited to images with a constant projection angle (e.g. 0°, corresponding to anterior-posterior projection), and predicts only 3D surface geometries rather than 3D volumetric configuration. To ensure fairness, we therefore conducted the comparison on this basis.

We trained our model as previously described, and using the data described in section 3.8.1. To train the IGCN model, we first extracted surface meshes from the ground-truth volumetric meshes in the training, validation, and test sets. We then trained the model using the deformed surface meshes and associated synthetic kV images for the

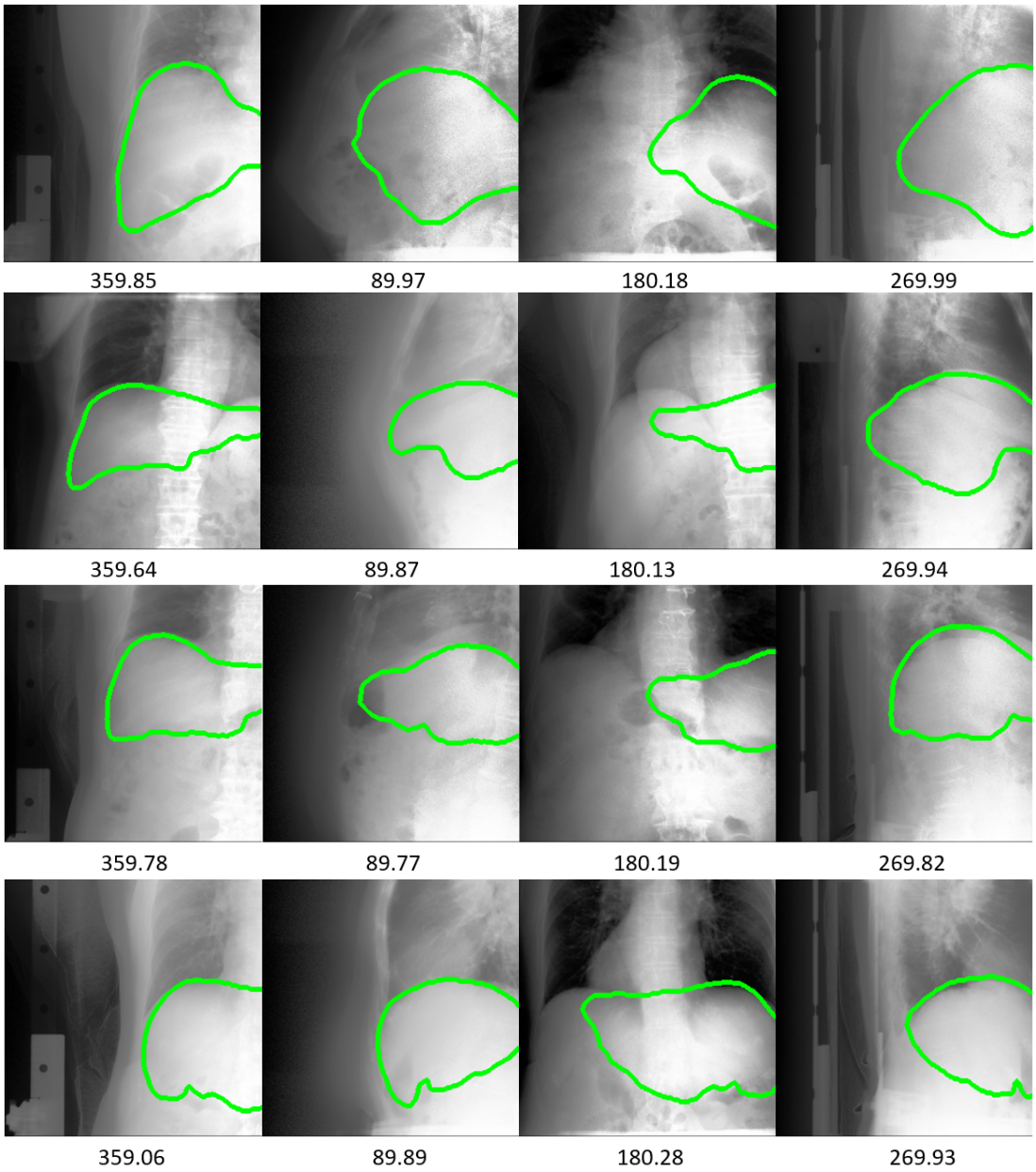


Figure 3.11: Samples of overlaid predicted liver boundary projections on corresponding real kV images for the four patients. Rows 1-4 show, respectively, results for patients 1-4. Results for images acquired at four projection angles (degrees, indicated below the images) are shown.

projection angle zero. Image dimensions were 256×256. Initial batch size, learning rate, and total number of epochs were as specified in [217]: 1, 0.0001 and 1000, re-

spectively.

Table 3.3: Summary statistics from performance comparison between our model and IGCN. All errors are computed with respect to predicted organ surface meshes. Patient case numbers are indicated in column 1. E_{Pred}^{Ours} and E_{Pred}^{IGCN} refer to prediction errors for our method and IGCN, respectively. U_{GT} refer to underlying ground-truth deformation magnitudes. *Mean (std)*: means (and standard deviations) of values across all nodes, all deformation states, and all projection angles. *Mean peak*: means of the peak values for each deformation state across all projection angles. *Max peak*: overall maximum values from all nodes, deformation states and angles. *99th Percentile*: 99$^{\text{th}}$ percentile values from all nodes, deformation states and angles. All values reported in mm.

Case		**Mean (std)**	**Mean peak**	**Max peak**	**99$^{\text{th}}$ Percentile**
1	E_{Pred}^{Ours}	0.17±0.11	0.89	4.91	0.47
	E_{Pred}^{IGCN}	0.18±0.25	1.13	6.37	0.93
	U_{GT}	10.11±1.24	13.94	28.12	13.53
2	E_{Pred}^{Ours}	0.14±0.15	1.53	7.13	0.67
	E_{Pred}^{IGCN}	0.17±0.21	2.81	8.79	1.18
	U_{GT}	11.37±1.55	15.41	34.91	14.09
3	E_{Pred}^{Ours}	0.15±0.13	1.46	13.83	0.77
	E_{Pred}^{IGCN}	0.19±0.23	2.05	14.31	1.23
	U_{GT}	14.52±2.39	18.71	49.63	16.74
4	E_{Pred}^{Ours}	0.12±0.09	0.78	3.98	0.44
	E_{Pred}^{IGCN}	0.14±0.17	0.97	5.31	0.77
	U_{GT}	10.01±1.05	12.17	25.64	11.33

Images with projection angle zero from the test set were passed to the IGCN model, which predicted corresponding deformed surface meshes. To facilitate a meaningful comparison with our model, therefore, we derived corresponding surface meshes from the latter's predicted volumetric meshes for the same images. The outcomes presented in Table 3.3 for the test set show that our method achieved overall higher accuracy in liver surface deformation across all patient cases. Further, we conducted one-way ANOVA tests for each test set to check for statistically significant differences in mean errors between our approach and IGCN. The resulting p-values of 0.0419, 0.0088, 0.0074, and 0.0485 for patients 1, 2, 3, and 4, respectively, confirmed this significance

(assuming 0.05 alpha value).

3.9 Ablation study

Table 3.4: Impact of reducing the number of image encoder MLPs. All values reported in mm.

Experiment	Mean (std)	Mean peak	Max peak	99th Percentile
1	0.21±0.18	1.92	7.53	1.19
2	0.23±0.22	2.88	8.47	1.45

This section describes the experiments we performed to determine the impact of each model component on the overall model performance. All experiments were conducted by training the model for 400 epochs. The synthetic data generated for patient 1 were used; hence all results should be compared with patient 1 values in Table 3.1).

We first assessed the impact of the four MLPs in the image encoder. As described, we incorporated an MLP for each convolutional layer, with each MLP responsible for five features per vertex, for a total of 20 features. Two experiments were conducted: 1) we removed MLPs associated with the first two convolutional layers, resulting in only ten image features per vertex; 2) we removed all MLPs except the final one, associated with the final convolutional layer, resulting in five features per vertex. In each of these cases, projection angle information was integrated into the input layer. Results are shown in Table 3.4, which demonstrates the importance of all four MLPs.

Table 3.5: Impact of removing gantry (projection angle) information from the image encoder. All values reported in mm.

Mean (std)	Mean peak	Max peak	99th Percentile
0.23±0.21	2.81	9.33	1.27

We next explored the impact of projection angle information in the image encoder by

omitting this information from the input layer. For this experiment, the number (i.e. four) of MLPs remained constant. The results in Table 3.5 show that feeding angle information to the input layer is more successful.

Table 3.6: Impact of removing skip connections from the mesh deformation network. All values reported in mm.

Mean (std)	**Mean peak**	**Max peak error**	99^{th} **Percentile**
0.20±0.19	2.04	7.69	1.31

In the next experiment, we focused on the effect of residual connections used in the mesh deformation network by simply removing them. Results in Table 3.6 indicate our approach with residual connections was superior.

Table 3.7: Impact of replacing graph-attention layers with graph convolutional network layers. All values reported in mm.

Mean (std)	**Mean peak**	**Max peak**	99^{th} **Percentile**
0.18±0.16	1.51	7.05	0.83

Finally, we explored the impact of the graph attention layers in the mesh deformation network by replacing them with GCN layers. The architecture was otherwise unchanged. Results in Table 3.7 demonstrate that utilizing GAT layers is more effective.

3.10 Summary

In this chapter, a novel patient-specific end-to-end deep learning approach is introduced. This method combines a CNN image encoder with a graph-attention network, incorporating learnable FPNs. The goal is to reconstruct 3D volumetric organ models from a single-view kV planar X-ray image with arbitrary gantry angles. To incorporate

information about arbitrary projection angles, an additional channel is added to the input image, facilitating the extraction of angle-dependent features. This design allows the model to reconstruct 3D anatomy from kV images acquired at any projection angle.

The fusion of image features into the 3D mesh space is achieved through four learnable FPNs. Each FPN is associated with its corresponding convolutional layer in the encoder, extracting hierarchical features for each vertex. This approach eliminates non-trainable components, such as the vertex projection operation, from the model architecture. Consequently, the model becomes end-to-end trainable, enhancing its overall effectiveness in reconstructing 3D volumetric organ models from single-view kV planar X-ray images with arbitrary gantry angles.

This approach possesses several appealing characteristics: it relies solely on readily available in-treatment imaging capabilities, avoiding the need for expensive and scarce systems like MRI; it eliminates the necessity for additional sensing to provide surrogate signals; and it does not require the implantation of invasive FMs. Furthermore, the model is end-to-end trainable, ensuring optimization of all components, particularly the image feature encoder, in terms of overall prediction accuracy. To the best of our knowledge, this represents the first example of a deep learning framework capable of accurately reconstructing volumetric 3D organ models from single-view images at arbitrary angles, thus enabling such reconstructions across the entire in-treatment scan series. We demonstrated the feasibility and accuracy of the technique through experimentation with data obtained from four liver cancer patients.

Chapter 4

An attention-based CNN framework for volumetric organ shape reconstruction from single-view 2D projections

4.1 Introduction

This chapter presents an efficient and accurate method that relies on a self-attention-based CNN framework for estimating 3D organ deformations from single in-treatment kV planar X-ray images. The motivation behind this work stems from the necessity to overcome the limitations of the Deep-Motion-Net, which was introduced in the preceding chapter. The main drawback of the approach is that organ mesh vertex displacements are predicted directly and independently of each other. That is, the model

outputs a $3N$ vector of vertex displacements, where N is the number of vertices. This impacts the requisite sizes of both the GNN itself (number of trainable parameters) and the training dataset (number of examples). It also ignores the obvious fact that, rather than being independent, the respiration-driven motions of points within the liver (or any other organ) are highly correlated. That is, the liver deforms not arbitrarily but according to identifiable patterns that may be well-described by an appropriately designed low-dimensional representation.

With this in mind, our present approach aims to solve the motion estimation problem by using image-derived features to predict latent parameters (with length $D \ll 3N$) from such a low-dimensional representation. The latter takes the form of a deep autoencoder (AE), pre-trained using synthetically generated organ motions. This significantly reduces the complexity of the estimation problem and enables a corresponding reduction in the number of learnable parameters. Training data and time are similarly reduced, while prediction accuracy is improved and remains negligibly dependent on image projection angle. The schematic diagram in Figure 4.1 provides a comprehensive overview of the workflow employed in our second approach.

In section4.2, we discuss the overview of the methodology. Section 4.3 illustrates model evaluation and results and the last section demonstrates the detailed description of the ablation study that has been conducted.

4.2 Methodology

In this section we present our deep learning approach, which is based on a two-step learning process: first, to estimate a low-dimensional representation of organ deforma-

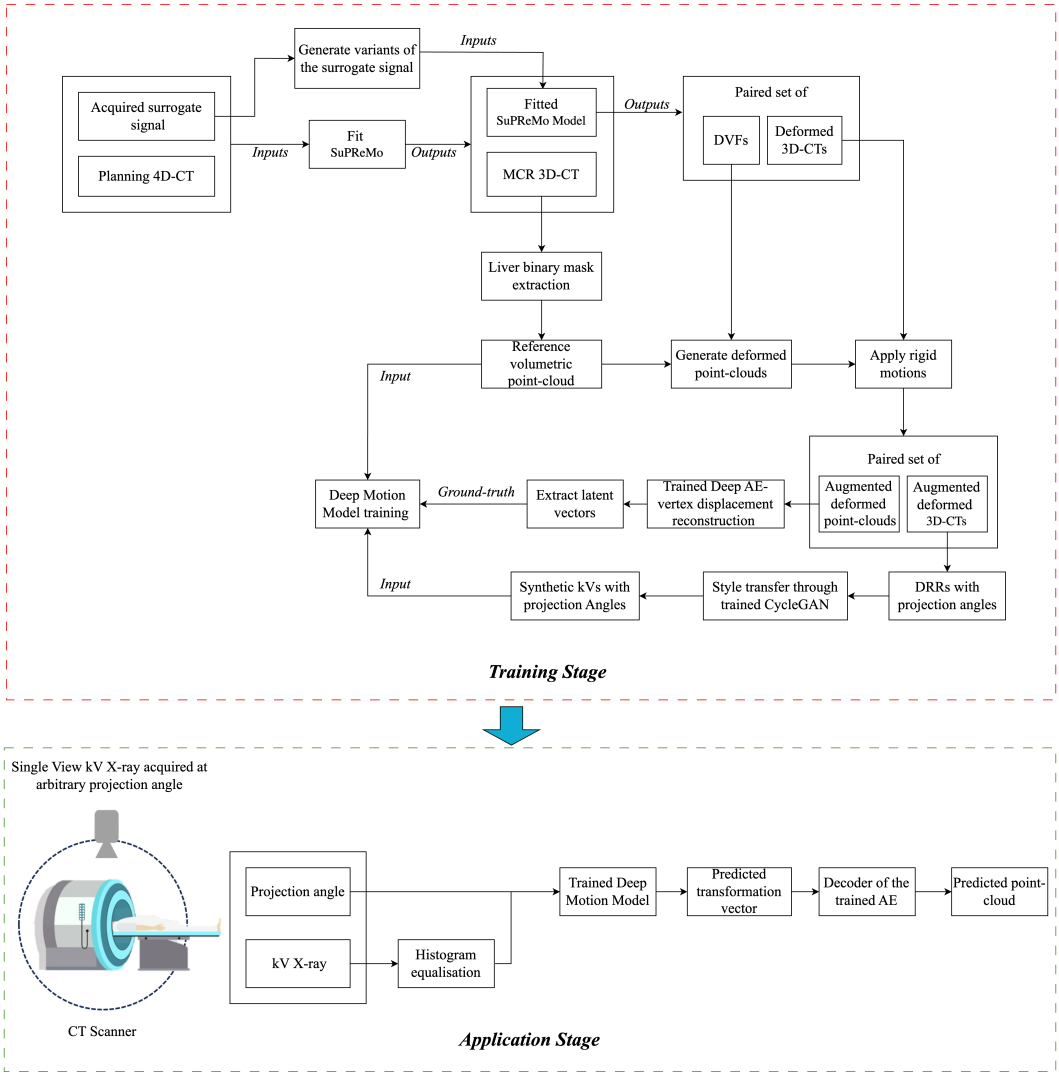


Figure 4.1: In the training stage, a synthetic motion dataset was created and a CNN model was trained to predict volumetric liver mesh deformation from a single X-ray projection. In the application stage, the trained CNN model was employed to derive the predicted deformed mesh using a real kV X-ray image captured at any arbitrary projection angle during the treatment process.

tions via a deep AE; and second, to learn a mapping between X-ray image features and the AE latent space parameters. The overall architecture is illustrated in Figure 4.2.

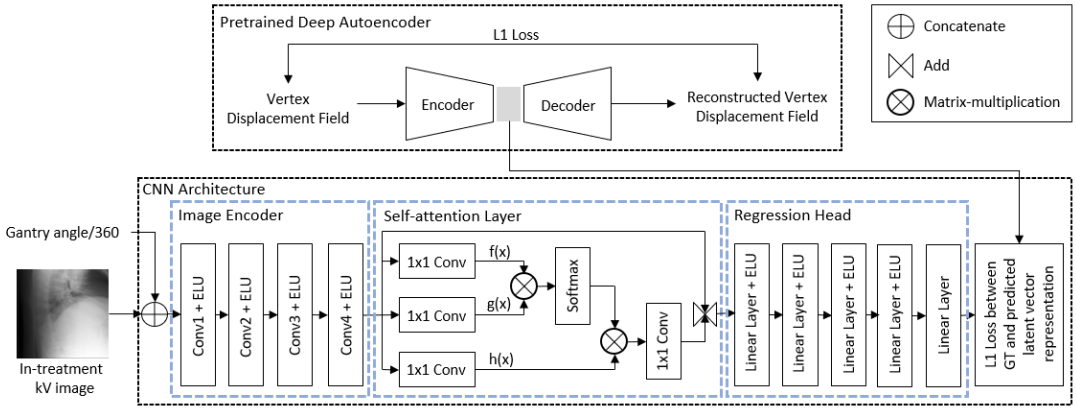


Figure 4.2: Illustration of the model architecture. A pre-trained deep autoencoder extracts latent vector representation of the vertex displacements for the deformed shape as ground-truth. A self-attention-based CNN predicts the low-dimensional representation by extracting projection angle-dependent semantic features from an input kV X-ray image.

4.2.1 3D organ shape representation

We use a sampled sparse point-cloud representation to describe the volumetric organ shape, rather than its surface only. The point-cloud can be described as a set of N vertices. Patient-specific organ template point-cloud, derived from reference CT volumes, are constructed for each patient.

4.2.2 Deep autoencoder for representing vertex deformations

As depicted in Figure 4.2, the AE operates on the $3N$-vectors of organ point-cloud vertex displacements. The encoder consists of two linear layers, with a non-linear activation (i.e. ELUs) for the first layer. Layers one and two comprise 20 and five neurons, respectively. The optimal latent space size was determined by analysing cumulative explained variance recovered from the training vertex displacement fields and utilizing a threshold of 0.995. The decoder also comprises two linear layers (with ELU activa-

tion following the first), with 20 and $3N$ neurons, respectively. As described in Section 3.6, the AE was trained on synthetically generated vertex displacement examples. The resulting latent representations for these examples were then used as ground-truth data for training the CNN model.

4.2.3 CNN for mapping image features to deformation parameters

Incorporating projection angle

The inputs to the attention-based 2D-CNN encoder are a single-view kV X-ray image and its corresponding projection angle, which varies from 0 to 360 degrees. We first normalize the projection angle by dividing it by 360. Since the projection angle is represented by a scalar value, the first step involves creating a 2D matrix that matches the dimensions of the input image. In this matrix, each element is assigned the scalar value corresponding to the projection angle. Subsequently, the resulting matrix is concatenated with the input image as an additional channel before it is sent to the image encoder.

Convolutional image encoder

Angle-dependant perceptual features of the input image are first extracted using four convolutional layers in the image encoder. These layers comprise 16, 32, 64, and 128 filters, respectively. Kernel size for all layers is set to 3×3, and stride and padding are each set to 1. ELU activations are applied for non-linearity after every convolutional operation. Output feature maps of each layer are down-sampled using 2×2 max-pooling layers with stride 2 before passing into the next layer.

Self-attention layer

We use a self-attention layer [244] after the last convolutional layer of the image encoder. It enables effective modelling of interactions between widely distant spatial regions, i.e. long-range, multi-level dependencies across image regions. We first perform 1×1 convolutions on the output feature map x of the last convolutional layer to produce new feature maps $f(x)$, $g(x)$ and $h(x)$ as shown in Figure 3.7. The rationale for using 1×1 convolution was to improve memory efficiency by lowering the number of channels: from C to C/k, where $k = 8$ [244]. We then perform matrix multiplication between $f(x)$ and $g(x)$ to acquire pairing covariances between all pixels and use a softmax layer to generate the attention map. Next, we generate the output feature map by multiplying the attention map with $h(x)$. Thereafter, we apply another 1×1 convolutional layer to render the depth/channels of the output feature map consistent with the number of channels in the input feature map. Subsequently, we scale the output feature map with a learnable parameter and add it to x as a residual connection. The scaling parameter is initialized to zero. This parameter allows the network to depend on local information first and progressively learn to allocate more weight to non-local evidence.

Regression head

The output from the self-attention layer is flattened and input to the first FC layer of the regression head. The regression head consists of five FC layers, with the first four consisting of 128, 64, 32, and 16 output channels respectively. For non-linearity, ELUs are used after each linear layer except the output layer. For the output linear layer, we set the number of output channels to five to obtain the predicted transformation vector.

In the image encoder and regression head, the output of all convolutional and FC layers, except the output layer, are normalised using Group Normalisation [225]. This helps to reduce the internal covariate shift, which regularly alters the distribution of the hidden-layer activations during model training.

4.2.4 Implementation and training

Both the AE and CNN framework were constructed using PyTorch. The Adam optimiser was employed to train both models, incorporating a weight decay of 0.001. Learning rates of 0.0001 and 0.0002 were used for the AE and CNN model, respectively. Batch size was consistently set at 16 for both training and validation datasets. To regulate the training process, early stopping was implemented, monitoring the validation loss for 30 consecutive epochs: if there was no improvement in the loss, the optimiser terminated the training. Additionally, if no progress was observed after eight consecutive epochs, the learning rate was decreased by a factor of 0.8. The gradient descent algorithm reached the optimal solution after 600 epochs. The entire training process took approximately eight hours using an Nvidia Quadro RTX 4000 GPU, and operating at a resolution of 256×256 pixels.

4.3 Model evaluation and results

The assessment of the model closely followed the evaluation methodology applied to the GNN model discussed in the preceding chapter, specifically concerning liver motion and utilizing clinical data from four liver cancer patients. Sections 4.3.1 and 4.3.2 outline the experimental setup and model evaluation based on synthetic respiratory motion scenarios. Furthermore, these sections conduct a comparative analysis of this

model with our previous model architecture described in Chapter 3. Subsequently, in Section 4.3.3, the qualitative evaluation of the model on in-treatment kV images for each patient case is presented. Furthermore, Section 4.3.4 conducts a comparative analysis of this model with the recently introduced IGCN model [217, 218], which predicts only 3D surface geometries using single kV X-ray images at projection angle zero. Section 4.4 showcases the ablation experiments that were conducted as part of the evaluation process.

4.3.1 Experiments on synthetic data derived from SuPReMo

Due to the absence of state-of-the-art methods for recovering volumetric mesh deformation using a single kV image with any projection angle, a comparison was made between our novel self-attention-based CNN approach and our prior graph-attention-based network method. Moreover, we explored an additional comparison by replacing the self-attention layer in the CNN model with a vision transformer [245] while keeping the image encoder and regression head unchanged. These three models were trained and validated under similar conditions, employing 256x256 image dimensions, with the only exception being the number of epochs. The two CNN-based networks required 600 epochs to reach the optimal solution, while the GNN network achieved this in 400 epochs.

Vision transformer configuration

The output feature maps of the final convolutional layer ($32 \times 32 \times 128$) from the image encoder were divided into 8×8 patches and each patch was embedded into a 128-dimensional vector. This process resulted in obtaining 16 patches for each image,

Table 4.1: Summary statistics for all test sets with synthetic kV images. Patient case numbers are indicated in column 1. E_{pred} and U_{GT} refer to prediction errors and underlying ground-truth deformation magnitudes, respectively. *Mean (std)*: means (and standard deviations) of values across all nodes, all deformation states, and all projection angles. *Mean peak*: means of the peak values for each deformation state across all projection angles. *Max peak*: overall maximum values from all nodes, deformation states and angles. *99th Percentile*: 99$^{\text{th}}$ percentile values from all nodes, deformation states and angles. All values reported in mm.

Case		**Mean (std)**	**Mean peak**	**Max peak**	**99$^{\text{th}}$ Percentile**
1	$E_{Pred}^{Ours-CNN}$	0.065±0.04	0.29	2.69	0.24
	$E_{Pred}^{Ours-GNN}$	0.16±0.13	1.39	6.75	0.75
	$E_{Pred}^{CNN-ViT}$	0.12±0.11	0.77	5.46	0.62
	U_{GT}	10.18±1.33	14.71	28.12	13.85
2	$E_{Pred}^{Ours-CNN}$	0.088±0.06	0.39	1.76	0.31
	$E_{Pred}^{Ours-GNN}$	0.18±0.19	1.99	7.97	0.97
	$E_{Pred}^{CNN-ViT}$	0.15±0.13	1.17	6.74	0.69
	U_{GT}	11.65±1.76	15.76	34.91	14.38
3	$E_{Pred}^{Ours-CNN}$	0.084±0.04	0.30	2.96	0.23
	$E_{Pred}^{Ours-GNN}$	0.22±0.34	3.29	14.66	1.81
	$E_{Pred}^{CNN-ViT}$	0.19±0.18	1.13	6.88	0.99
	U_{GT}	14.89±2.54	19.36	49.63	17.65
4	$E_{Pred}^{Ours-CNN}$	0.059±0.04	0.25	1.36	0.19
	$E_{Pred}^{Ours-GNN}$	0.12±0.11	1.16	4.36	0.51
	$E_{Pred}^{CNN-ViT}$	0.093±0.10	0.65	4.00	0.55
	U_{GT}	10.07±1.13	12.86	25.64	11.97

with each patch represented as a [1×128] vector. Next, the regression token embedding vector was prepended to these learnable embedding vectors, followed by incorporating positional embeddings into all of these vectors. The positional encoding enables the model to understand the spatial positioning of each patch within the input image. The regression token was positioned as the first token in each sequence, serving as input to the regression head (final MLP network).

Subsequently, we passed the patch embeddings along with the regression token through two transformer blocks, each consisting of a Multi-Head Self-Attention (MSA) Block

and an MLP Block. Within the MSA layer, each input patch embedding was linearly transformed into three distinct vectors: query, key, and value. The MSA layer then computed the dot product between the query vector of each patch and all key vectors, scaling the result by the square root of the vector dimensionality. The resulting scores, known as attention weights, were normalized using the softmax function to obtain attention coefficients. These attention coefficients were used to weight and combine the corresponding value vectors associated with the key vectors, yielding the final output of the attention mechanism.

For the MLP block, we employed three linear layers with output channels of 256, 256, and 128, respectively. Each linear layer was followed by the Exponential Linear Unit (ELU) activation function, except for the output layer.

Finally, we extracted the regression token from the output vector of the Transformer blocks and passed it through the regression head to obtain the predicted transformation vector. Due to memory constraints, we limited the number of self-attention layers to two. Additionally, a dropout rate of 0.3 was applied during model training to mitigate overfitting.

Comparison with test dataset

The results for the test datasets of all four patients are presented in Table 3.1, with Euclidean distance used as the metric to evaluate distance errors between ground-truth and estimated shapes. Across patients, the CNN with self-attention model consistently outperformed both the GNN model and CNN with vision transformer model (CNN-ViT) in terms of mean errors, achieving values ranging from 0.059±0.04 mm to 0.088±0.06 mm. In contrast, the GNN's mean errors ranged from 0.12±0.11 mm to

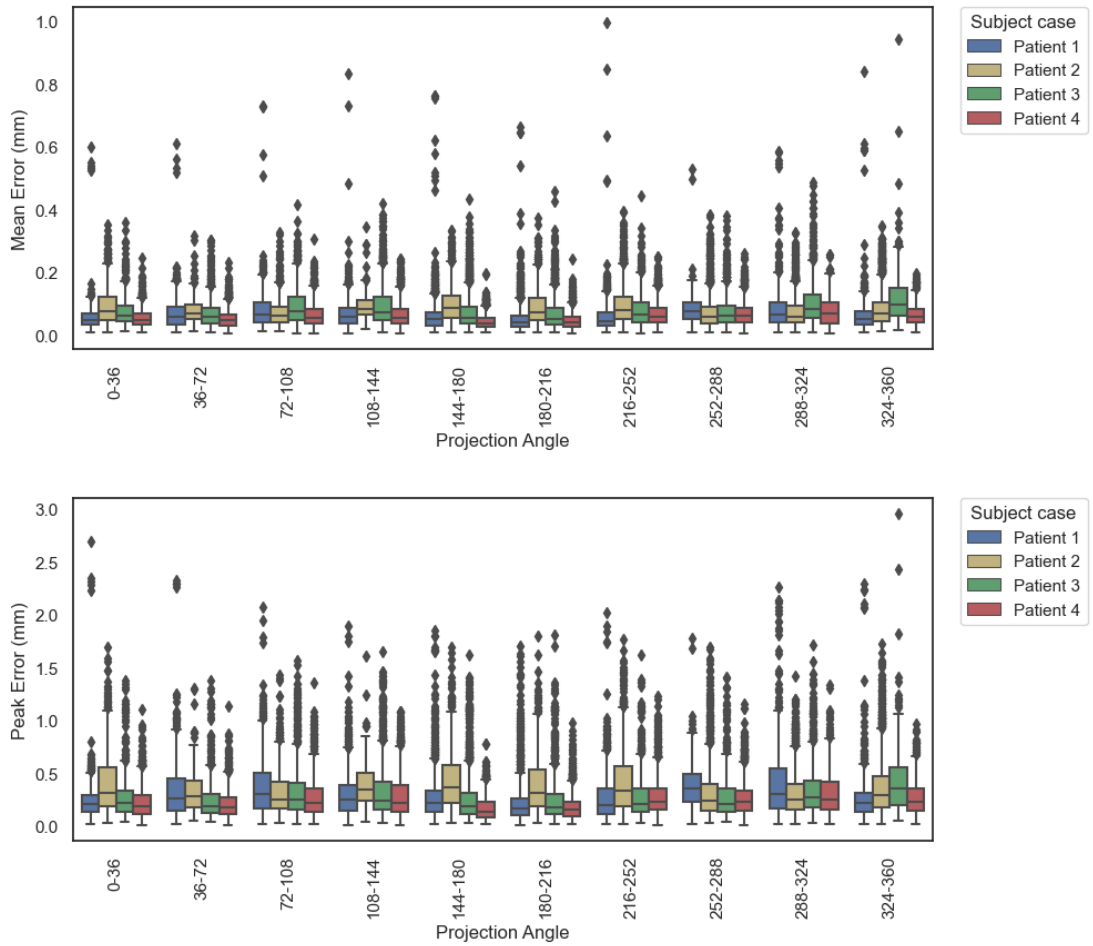


Figure 4.3: Effect of projection angle on prediction accuracy: box and whisker plots of the mean (top) and peak (bottom) prediction errors grouped according to image projection angle (degrees). Each box and whisker shows the distribution of errors for the indicated projection angle using all deformation states in the test set. For clarity of visualisation, angles are further grouped into 10 equal bins covering a full revolution. Results for patients 1 (blue), 2 (yellow), 3 (green), and 4 (red) are shown for each bin.

0.22±0.34 mm, and CNN-ViT's mean errors ranged from 0.093±0.10 mm to 0.19±0.18 mm.

Analyzing mean peak errors, the CNN with self-attention model consistently displayed lower values (0.25 mm to 0.39 mm) compared to the GNN model, which exhibited higher mean peak errors ranging from 1.16 mm to 3.29 mm, and the CNN-ViT model,

which demonstrated mean peak errors ranging from 0.65 mm to 1.17 mm. Similarly, for max peak errors, the CNN with self-attention model consistently showcased smaller values (1.36 mm to 2.69 mm) compared to the GNN model (4.36 mm to 14.66 mm) and CNN-ViT model (4 mm to 6.88 mm).

Regarding the 99^{th} percentile, the CNN with self-attention model yielded smaller values (0.19 mm to 0.31 mm) compared to the GNN model with higher 99^{th} percentile values (0.51 mm to 1.81 mm) and CNN-ViT model with higher 99^{th} percentile values (0.55 mm to 0.99 mm). These findings collectively indicate that the CNN with self-attention model consistently outperforms the GNN model and CNN-ViT model across all error metrics for liver deformation estimation in the specified patient datasets.

Patient 3 exhibits the highest errors among all four cases for both models. The predominant factor contributing to the elevated errors in this patient case is the substantial presence of reconstruction artefacts within the 4D-CT data. More specifically, it is observed that in this patient's scenario, the deformed 3D-CT image volumes and the corresponding DVFs, particularly at the initial and final time points, comprise significant errors. This circumstance renders it considerably challenging for the model to generate accurate predictions.

To further validate these findings, one-way ANOVA tests were conducted on each test dataset to assess potential statistically significant differences in mean error values. To this end, we compared the mean error values of our self-attention-based CNN approach with our GNN-based method, as well as those of our self-attention-based CNN model with the CNN-ViT model. The corresponding P-values were calculated, and they are all extremely small, with nearly all of them approximating zero for all patients. These results consistently demonstrate that all P-values were less than the 0.05 significance

level, providing strong evidence that the mean errors of the self-attention-based CNN approach are significantly smaller than those obtained with the GNN-based approach and the CNN-ViT model.

Figure 4.3 depicts the visualization of the variability of the mean and peak error against the projection angle for the test datasets with box and whisker plots for all four liver patient cases. By looking at this figure, we can see the estimated deformations are independent of the projection angle because there is no considerable difference between the distribution of the bars and they are almost lined. Finally, it is crucial to highlight that the elevated peak errors observed in all instances were highly localized within the point-clouds. This is evidenced initially by the consistently low mean values and the 99^{th} percentile errors (refer to Table 4.1), which remained below 1 mm for all test sets. This is further demonstrated through the visual representations of predicted shapes, where displacement errors are colour-mapped, presented in the right columns of Figure 4.4. The errors are minimal (≤ 2 mm) across most of the surface, with only isolated regions exhibiting higher values.

4.3.2 Experiments on synthetic data derived from 4D-Precise

We have conducted a comparative analysis to assess the accuracy of our trained model against an alternative synthetic dataset, which was generated through a distinct methodology from SuPReMo-based synthetic data. This new synthetic dataset comprised motion scenarios derived from 4D-Precise [246], a self-supervised deep-leaning model that generates deformation fields by leveraging real kV X-ray images as input. The 4D-Precise model exhibits the capability to estimate voxel-wise motion fields while concurrently reconstructing a 3D-CT volume for any arbitrary time point within the

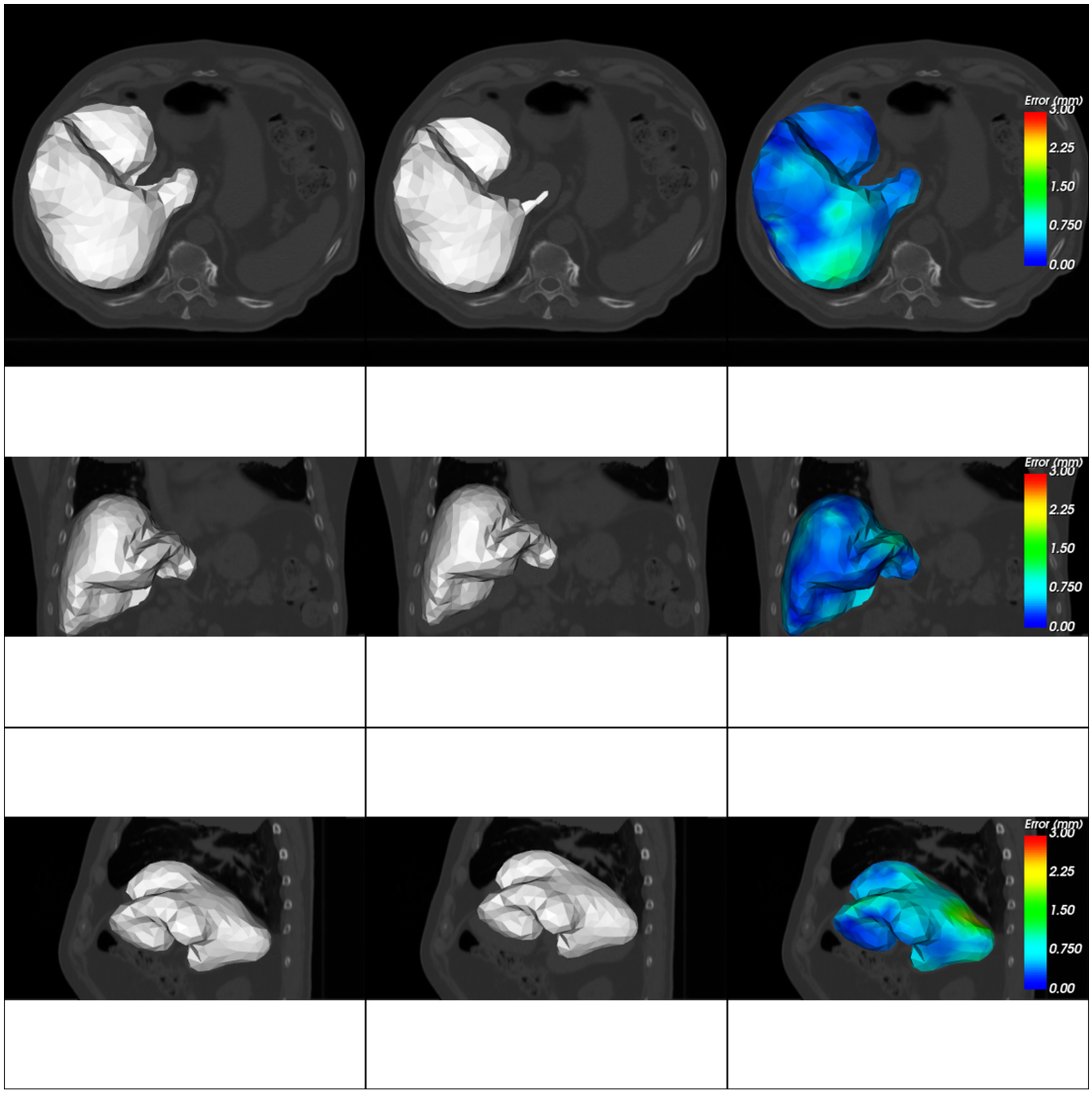


Figure 4.4: Visualisations of ground-truth deformed (left column), template (middle column), and estimated deformed (right column) 3D liver shapes. Point-clouds as mesh representations are overlaid on the deformed 3D-CT volume. Rows 1-3 show, respectively, axial, coronal, and sagittal views. Results are shown for the *worst* performing test case for patient 1: image projection angle 23.679°, and deformation state producing highest errors. Contours in the right column indicate the spatial distribution of errors on the surface. Similar results for patients, 2-4 are presented in Appendix B.

input kV X-ray projections.

The training phase of the 4D-Precise model involved in-treatment kV images specific to each patient case, utilizing the same reference 3D-CT image volume employed in SuPReMo for generating synthetic motion instances. Once the 4D-Precise model was adequately trained, we proceeded to extract the displacement field for each real kV image from the second scan series (see Section 4.3.3) of each patient case. The subsequent interpolation of the displacement vector field at the positions of the reference point-cloud representation allowed us to produce deformed point-cloud configurations. These synthetically generated point clouds were subjected to a comparison with the corresponding predicted point cloud representations generated by our model. In this experimental setup, we utilized uniformly sampled 135, 132, 132, and 131 input kV images for patients 1, 2, 3, and 4, respectively.

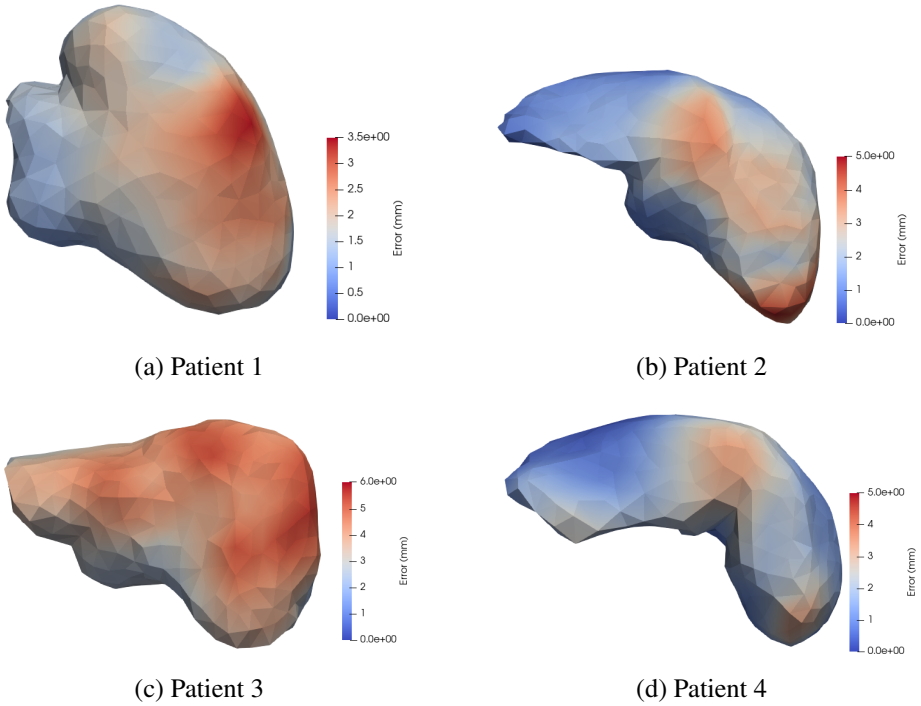
Table 4.2 illustrates the summary statistics for the discrepancies between 4D-Precise and our model predictions for all four patient cases. The mean differences for the CNN model span from 0.87 mm to 1.30 mm, with patient 4 exhibiting the lowest and patient 3 the highest values, whereas for the GNN model, the mean differences range from 1.13 mm to 2.08 mm, with patient 4 having the minimum and patient 3 the maximum. In mean peak evaluation, patient 1 has the lowest mean peak discrepancy for both CNN (2.09 mm) and GNN (3.27 mm) models, whereas patient 3 demonstrates the highest mean peak difference for both models, with 3.01 mm for CNN and 4.55 mm for GNN. Regarding max peak, patient 1 consistently displays the lowest values for both CNN (3.44 mm) and GNN (5.01 mm), whereas patient 3 consistently exhibits the highest max peak differences, reaching 5.61 mm for CNN and 8.47 mm for GNN. For the 99^{th} percentile values, patient 1 shows the lowest (1.87 mm) for CNN and (2.85

Table 4.2: Summary statistics for all test sets with real kV images produced from 4D-Precise. Patient case numbers are indicated in column 1. E_{pred} refer to prediction discrepancies between our model and 4D-Precise whereas $U_{SD-4DPrecise}$ refer to the underlying synthetic deformation magnitudes of the point-clouds generated using 4D-Precise. *Mean (std)*: means (and standard deviations) of values across all nodes, and all deformation states. *Mean peak*: means of the peak values for each deformation state. *Max peak*: overall maximum values from all nodes and deformation states. *99th Percentile*: 99$^{\text{th}}$ percentile values from all nodes and deformation states. All values reported in mm.

Case		**Mean (std)**	**Mean peak**	**Max peak**	**99$^{\text{th}}$ Percentile**
1	$E_{Pred}^{Ours-CNN}$	1.10±0.26	2.09	3.44	1.87
	$E_{Pred}^{Ours-GNN}$	1.62±0.41	3.27	5.01	2.85
	$U_{SD-4DPrecise}$	1.98±1.84	10.41	18.83	8.94
2	$E_{Pred}^{Ours-CNN}$	0.89±0.52	2.82	4.86	2.34
	$E_{Pred}^{Ours-GNN}$	1.32±0.71	4.55	6.81	3.73
	$U_{SD-4DPrecise}$	3.89±2.75	18.10	25.44	13.71
3	$E_{Pred}^{Ours-CNN}$	1.30±0.40	3.01	5.61	2.54
	$E_{Pred}^{Ours-GNN}$	2.08±0.67	4.43	8.47	4.06
	$U_{SD-4DPrecise}$	4.71±2.20	13.82	24.78	12.01
4	$E_{Pred}^{Ours-CNN}$	0.87±0.44	2.75	4.76	2.33
	$E_{Pred}^{Ours-GNN}$	1.13±0.59	3.66	6.49	3.09
	$U_{SD-4DPrecise}$	3.83±2.79	14.24	24.96	12.77

mm) for GNN, while patient 3 exhibits the highest values, with 2.54 mm for CNN and 4.06 mm for GNN. Overall, patient 3 consistently exhibits higher metrics for both CNN and GNN models, while patient 4 tends to have lower values for both models. The CNN model consistently outperforms the GNN model across all patients based on the discrepancies calculated relative to the liver deformation derived from 4D-Precise motion fields.

To validate these findings, one-way ANOVA tests were performed on each test dataset, assessing any statistically significant differences in mean values between our GNN-based and CNN-based approaches. The calculated P-values, which consistently came towards zero for all patients, were discovered to be extremely small. These results



(a) Patient 1

(b) Patient 2

(c) Patient 3

(d) Patient 4

Figure 4.5: The spatial distribution of displacement discrepancies on the surfaces of the worst-performing test cases is depicted, illustrating the differences between 4D-Precise and our CNN model predictions using the same input real kV images.

consistently show that all P-values are less than the 0.05 significance level, implying that the mean differences of the self-attention-based CNN approach are significantly smaller than those obtained with the GNN-based method.

Figure 4.5 presents a visualization of the displacement discrepancies (only for the worst performing synthetic test cases) between 4D-Precise and our CNN model predictions using the same input real kV images. However, determining the more accurate prediction is challenging due to the absence of true ground-truth information associated with each real kV image. The disparities are clearly higher compared to SuPReMo-generated deformations. In the case of liver shapes, the peak displacement differences are spatially concentrated over the surface, resulting in a lack of highly

localized regions.

4.3.3 Evaluation on real kV images

In our second set of experiments, we utilized real in-treatment kV images from the second scan series of each patient as described in the previous Chapter. Since ground-truth deformations were unavailable, we employed two evaluation approaches: 1) a semi-quantitative assessment based on an image similarity metric between input real kV images and model-generated DRRs; and 2) a qualitative assessment relying on overlaying model-predicted liver boundaries on input kV images. For the qualitative assessment, we considered all images in the scan series. To expedite computation time, especially during spline deformation of the image volumes, we sampled 100 images uniformly from each patient's series for the similarity-based assessment.

Mutual Information-based assessment

In this section, we used the same systematic approach outlined in Figure 3.10, which we employed to assess real kV images in the absence of ground-truth data.

Table 4.3: MI similarity scores (mean ± standard deviation, computed from the 100 images sampled from each scan series) between real kV images and DRRs generated at the same projection angles for each patient case. Column 2 presents values when the reference (i.e. undeformed) CT volume is used. Column 3 presents values when the CT volume is deformed using the model-predicted deformation fields. All values were computed on the liver region.

Case	Reference	Deformed
1	1.16±0.31	1.33±0.21
2	1.14±0.21	1.40±0.12
3	1.13±0.27	1.30±0.21
4	1.31±0.25	1.45±0.13

Table 4.3 summarizes the results of these experiments. The similarity score of the reference configuration is lower than the motion-corrected configuration indicating that our model improves the motion of the liver. Since we assessed 100 projection angles, the values in Table 4.3 were calculated by averaging each similarity metric value across all 100 kV images per scan series. Moreover, the marginal disparities in average MI values led us to conduct a one-way ANOVA test for each patient case. The resulting P-values for MI differences between the reference and deformed versions were 0.0365, 0.0259, 0.0438, and 0.0427 for patients 1, 2, 3, and 4, respectively, suggesting statistically significant (assuming alpha value of 0.05) differences.

Qualitative assessment by boundary overlay

As the initial step, the trained models were utilized to estimate deformed 3D shapes by inputting all histogram-equalized real kV images from the second scan series. Subsequently, each predicted shape generated a corresponding binary image volume. The liver surface boundaries projected onto the images were derived through ray-tracing on these binary volumes. These projected binary masks were then superimposed onto the corresponding input kV images. Figure 4.6 presents examples from each patient, and additional animated versions covering the entire scan series for each patient case are provided in the supplementary materials for a more comprehensive examination. These animated versions effectively demonstrate the qualitative consistency of the model predictions with the input images across multiple breathing cycles and the complete treatment gantry rotation.

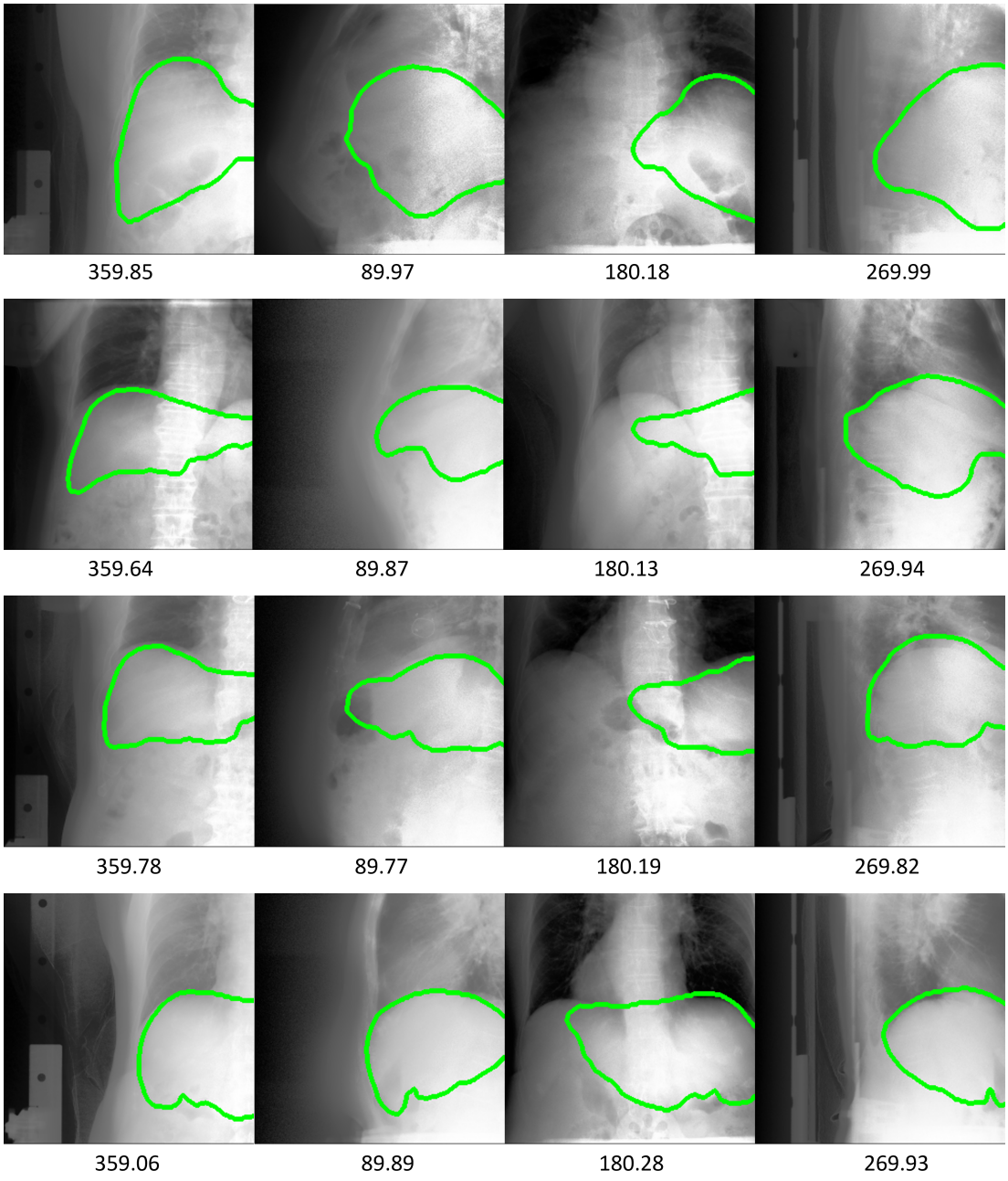


Figure 4.6: Samples of overlaid predicted liver boundary projections on corresponding real kV images for the four patients. Rows 1-4 show, respectively, results for patients 1-4. Results for images acquired at four projection angles (degrees, indicated below the images) are shown.

4.3.4 Comparison model with liver surfaces

Since there are no state-of-the-art techniques available for reconstructing volumetric point-clouds from arbitrary projection angles, methods do exist for reconstructing surface meshes from particular projection angles [210, 212, 215, 217, 218, 220]. In this section, we provide a comparison between our CNN-based approach and the IGCN [217, 218] model, which focuses on reconstructing liver surface mesh deformation from front-view kV projections, specifically at projection angle zero. This evaluation was conducted utilizing test datasets derived from SuPReMo. To ensure a comprehensive evaluation, we used the exact same approach as mentioned in Section 3.8.3 in the previous Chapter.

The results, as presented in Table 4.4, indicate that our latest CNN-based method consistently achieved higher accuracy in modelling liver surface deformation across all patient cases in the test set. Furthermore, we performed one-way ANOVA tests on each test dataset to assess whether statistically significant differences in mean errors between our CNN approach and IGCN. The computed P-values are extremely small values for all patients (assuming an alpha value of 0.05). This further indicates that the mean errors of our CNN-based approach are significantly smaller than IGCN.

4.4 Ablation study

This section exemplifies the ablation experiments we performed to determine the impact of self-attention and projection angle components on the overall model performance. All these experiments were performed utilizing test datasets derived from SuPReMo, and the model was trained for 600 epochs. The results should be compared

Table 4.4: Summary statistics from performance comparison between our CNN-model, GNN-model, and IGCN. All errors are computed with respect to predicted organ surface shapes. Patient case numbers are indicated in column 1. E_{Pred}^{Ours} and E_{Pred}^{IGCN} refer to prediction errors for our self-attention-based CNN method and IGCN, respectively. U_{GT} refer to underlying ground-truth deformation magnitudes. *Mean (std)*: means (and standard deviations) of values across all nodes, all deformation states, and all projection angles. *Mean peak*: means of the peak values for each deformation state across all projection angles. *Max peak*: overall maximum values from all nodes, deformation states and angles. *99th Percentile*: 99$^{\text{th}}$ percentile values from all nodes, deformation states and angles. All values reported in mm.

Case		**Mean (std)**	**Mean peak**	**Max peak**	99$^{\text{th}}$ **Percentile**
1	E_{Pred}^{Ours}	0.066±0.03	0.20	2.20	0.18
	E_{Pred}^{IGCN}	0.18±0.25	1.13	6.37	0.93
	U_{GT}	10.11±1.24	13.94	28.12	13.53
2	E_{Pred}^{Ours}	0.067±0.05	0.31	1.18	0.25
	E_{Pred}^{IGCN}	0.17±0.21	2.81	8.79	1.18
	U_{GT}	11.37±1.55	15.41	34.91	14.09
3	E_{Pred}^{Ours}	0.092±0.05	0.32	2.47	0.26
	E_{Pred}^{IGCN}	0.19±0.23	2.05	14.31	1.23
	U_{GT}	14.52±2.39	18.71	49.63	16.74
4	E_{Pred}^{Ours}	0.073±0.04	0.26	1.19	0.22
	E_{Pred}^{IGCN}	0.14±0.17	0.97	5.31	0.77
	U_{GT}	10.01±1.05	12.17	25.64	11.33

with the values for patient 1 in Table 4.1) since we utilized synthetic data generated for this specific patient.

Table 4.5: Impact of removing self-attention layer. All values reported in mm.

Mean (std)	**Mean peak**	**Max peak**	99$^{\text{th}}$ **Percentile**
0.081±0.13	0.61	4.64	0.55

We first compared the impact of the self-attention layer on model performance by removing this module. The architecture was otherwise unchanged. Results in Table 4.5 demonstrate that utilizing the self-attention layer is more effective.

In the subsequent experiment, we investigate the influence of excluding projection an-

Table 4.6: Impact of removing gantry (projection angle) information from the input layer. All values reported in mm.

Mean (std)	**Mean peak**	**Max peak**	99^{th} **Percentile**
0.076±0.09	0.48	5.97	0.43

gle information from the CNN model's input layer while keeping the other components constant. The outcomes presented in Table 4.6 demonstrate that including angle information in the input layer is more effective.

Table 4.7: Impact of FC layers in the regression head. All values reported in mm.

Experiment	**Mean (std)**	**Mean peak**	**Max peak**	99^{th} **Percentile**
1	0.079±0.08	0.54	3.82	0.41
2	0.14±0.12	0.72	5.19	0.67

Two experiments were conducted to assess the impact of FC layers in the regression head for the first patient case. In the first experiment, the last FC layer was removed, resulting in a new regression head with four FC layers. The first three layers had 128, 64, and 32 output channels, respectively, and the output linear layer consisted of five neurons to obtain the predicted transform vector.

In the second experiment, the last two FC layers were removed, leading to a regression head with three FC layers. The first two layers had 128 and 64 output channels, respectively, and the output layer consisted of five neurons for predicting the transform vector. The results of these experiments, as displayed in Table 4.7, emphasize the importance of utilizing all five FC layers in the regression head for this task.

Throughout these two experiments, all other components were held constant and unchanged.

4.5 Summary

In this chapter, a novel predictive framework, leveraging self-attention within a CNN, has been introduced. This framework facilitates for estimating 3D organ deformations from single in-treatment kV planar X-ray images captured at any arbitrary projection angle. The work is motivated by the need for accurate characterisation of patient anatomical motion during radiotherapy, to enable treatment adaptation. The proposed approach combines: 1) a deep AE that first learns low-dimensional representations of patient organ deformations; and 2) a self-attention-based CNN that learns mappings between deep semantic X-ray image features and corresponding encoded deformation latent representations. Learnt image features, moreover, are angle-dependent, meaning input X-ray images may be acquired at arbitrary projection angles. Full organ deformation fields are subsequently reconstructed by passing the predicted latent vectors to the AE decoder network. Since only low-dimensional latent vectors, rather than full displacement fields, are directly predicted, the network size and training data requirements are relatively small.

The motivation for introducing this new model architecture stems from the computational efficiency considerations, particularly in comparison to the more time-consuming GNN model, which is described in Chapter 3, employed in the training process. This approach exhibits enhanced robustness during inference for each input image, requiring only approximately 4 milliseconds per input image, whereas the GNN method demands around 27 milliseconds.

This framework not only addresses the efficiency concerns associated with the GNN model but also presents a promising advancement in accurately recovering 3D anatomical deformations from single kV planar X-ray images. Utilizing self-attention within

the CNN contributes to the model's ability to capture complex relationships in the data, resulting in improved efficiency and reduced inference time.

Chapter 5

Discussion & Conclusion

5.1 Contribution and summary

The thesis presented two novel patient-specific deep motion models for recovering 3D volumetric organ shape deformation from a single in-treatment kV planar X-ray image acquired at arbitrary projection angles. These two approaches have several attractive features: they use only readily accessible in-treatment imaging capabilities, rather than expensive and rare systems like MRI; they require no extra sensing to provide surrogate signals; and no invasive FM implantation. To the best of our knowledge, this is the first example of deep learning frameworks able to reconstruct volumetric 3D organ models accurately from arbitrary-angled single-view images, and thereby enable such reconstructions across complete in-treatment scan series. The predictive performance and the feasibility of the proposed networks were evaluated by using data from four liver cancer patients, with a focus on liver motion.

The first model, which employs a GNN framework as described in Chapter 3, requires

more training time. Therefore, an alternative approach was introduced, as described in Chapter 4, which focuses on predicting transformation parameters using a simple CNN network for vertex displacement prediction. This approach allows for a substantial reduction in the number of trainable model parameters, leading to fewer transformation parameters in the final prediction. Consequently, the CNN-based method demonstrates a faster processing time per input image during inference compared to the GNN method.

As mentioned, our two approaches have been developed specifically to accommodate input kV images acquired at arbitrary projection angles. We demonstrated this by training and evaluating the models with images generated at different projection angles. As shown in Figure 3.8 and Figure 4.3, the prediction accuracy was indeed virtually independent of projection angle.

The models were trained with synthetic respiratory motion data constructed using the SuPReMo toolkit. Training in this way is essential in the absence of ground-truth deformations corresponding to real kV images; that is, there appears to be no other way of acquiring paired deformation/image sets for this scenario. For similar reasons, direct quantitative evaluation of the model performance was also carried out using synthetic data. Naturally, the performance of the model will depend on the fidelity with which real patient motions are reproduced in the synthetic data. The evaluation process has been conducted in two directions where in the first case, the models were tested quantitatively on synthetic respiratory motion scenarios and qualitatively on in-treatment images acquired over a full scan series for liver cancer patients.

A key part of the synthetic data creation was the development of a method for generating realistic synthetic kV X-ray images. While DRRs produced from ray tracing

through CT volumes ultimately are also X-ray-based images, they do not suffer from the same scatter and noise phenomena of in-treatment kV X-ray images. Their appearance, consequently, is noticeably different. Therefore we first trained a CycleGAN [231] for each patient by conditioning on projection angle to learn the genuine kV X-ray style that can be transferred to the DRRs. CycleGANs are particularly well-suited to this task since they require only unpaired sets of DRRs and kV images. Initially, we utilized transposed convolution layers for the decoder portion of the generators; however, we encountered checkerboard-like artefacts in several instances. To address this issue, we transitioned to using convolution layers followed by pixel shuffle layers [234]. Pixel Shuffle effectively mitigates checkerboard artefacts by reorganizing and rearranging feature map elements to enhance spatial resolution without introducing interpolation-related blurriness. By reshuffling and aggregating elements within each channel of the feature map, Pixel Shuffle ensures that neighbouring pixels in the output correspond to adjacent regions in the input, thus maintaining spatial coherence and alignment. This structured reordering of elements helps alleviate irregular pixel arrangements that can lead to checkerboard patterns, ultimately contributing to the production of high-quality, artefact-free images in GANs.

We observed that the model prediction accuracy can be influenced by the quality and characteristics of the CT volumes from which training data are constructed. The higher peak errors were found for patient 3. The CT volumes of this patient case exhibited certain reconstruction artefacts related to motion. For this patient, in large amplitude deformed cases, these resulted in some mesh-image misalignment problems even in the ground-truth data (specifically, with the deformed states at initial and final time points) produced from SuPReMo. The deformed images for Patients 1 and 4, by contrast,

contained no obvious artefacts or ambiguous anatomy, and peak errors were correspondingly smaller (Table 3.1 and Table 4.1). As emphasised, however, the regions of higher peak errors, even in patient 3, were nonetheless very localised, and mean errors remained low.

The displacement discrepancies show clearly a quite higher value between 4D-Precise and our model when compared to the results obtained with SuReMo-based synthetic test datasets. In our analysis, we employed 4D-Precise as an independent approach to predict the output for uniformly sampled real kV images in the second scan series, aiming to compare these results with the predictions from our model. More importantly, the predictions from 4D-Precise may contain errors, as it relies on a self-supervised deep learning technique capable of extracting DVFs as the output to create DRRs similar to real kV X-rays. Although we utilized the motions estimated by 4D-Precise from real kVs, they do not represent the ground-truth concerning the true motions observed in the kV X-ray images.

In summary, the self-attention-based CNN model demonstrated superior performance over the GNN model in predicting 3D organ deformations from single in-treatment kV planar X-ray images. This was evidenced by lower overall error values and a higher precision in predicting vertex displacements.

5.2 Limitations and future directions

This thesis focuses on respiratory motion estimation, however, our techniques can in principle be used to predict any motion patterns, e.g. the peristaltic motion of the gut or longer-term structural changes, given appropriate training data. These models

involve no assumptions of periodicity or other specific motion characteristics. Clearly, the generation of training data is more challenging in some scenarios than in others, but this presents a practical difficulty rather than a theoretical one.

While SuPReMo appears to do a reasonable job of characterising the motion present in the input 4D-CT data, these data represent only averaged breathing cycles and do not by themselves provide information about the variability of this motion. Pragmatically, therefore, in this work, we assumed some bounds on the variations from this average and randomly generated motion states within these (Sect. 3.6.2). We will explore more rigorous approaches to characterising the patient motion variability (and incorporating this in model training data) based on kV image sequences acquired over several minutes during treatment.

Another shortcoming of our methods, as currently implemented, is that we only estimate the deformation of the target organ; nearby OARs and other anatomy are ignored. However, the model itself imposes no restrictions in this respect. If suitable training data covering the relevant anatomy, and meshes of the relevant anatomical structures can be generated, the model can in principle estimate motions for these in the same way. SuPReMo, for example, provides DVFs for the whole image volume, which could be used for this purpose. OAR and other meshes can normally be created similarly to the liver meshes used here. Potentially, the prediction of deformations for multiple disconnected meshes could result in anomalous overlapping regions. However, such instances would hopefully be rare when the training data include no such behaviour. If needed, separate non-overlap constraints in the formulation could also be conceived, which penalise mesh interpenetration in a way analogous to some contact formulations in computational mechanics. Albeit, this would increase the complexity

of the approach. A simpler alternative could be deploying a single mesh covering all relevant anatomy. In all scenarios, it is possible that both model and training data size requirements would increase, though the overall approach would remain the same.

As mentioned, two general approaches to adapting RT treatments could be considered using motion estimates generated by our technique: inter- and intra-fraction adaption. Inter-fraction plan adaption could be achieved by retrospectively estimating the motion that occurred during a treatment fraction and computing motion-compensated spatial distributions of delivered dose (dose accumulation); the treatment plan could then be adapted accordingly for subsequent fractions. No new technology seems to be needed for this, beyond the requirement to image throughout the fraction, rather than during initial patient positioning only. This would enable re-planning of subsequent fractions to mitigate the effects of unexpected motion, lowering toxicity and improving patient outcomes. To this end, a Clinical Scientist, Marcus Tyyger, at Leeds NHS Trust, is currently working on estimating motion-compensated dose accumulation by utilizing the predicted liver geometries for a full kV scan series generated from our self-attention-based CNN technique. The predicted deformed volumetric shape for each real kV image in the acquired scan series provides a snapshot of the patient's 3D anatomy at the specific time point. The process is as follows: 1.) Converting the CNN model output, in mesh format, into a format suitable for RayStation. This necessitates generating the predicted (deformed) CT for each kV X-ray projection using an image registration technique (e.g. Thin-plate-spline transformation); 2.) the positions of the gantry/MLCs at the start and end of the time period are determined, with interpolation between them; 3.) dose for the predicted 3D-CT anatomy is computed; 4.) deformable image registration is conducted between the reference 3D-CT and the predicted 3D-CT, and 5.)

the dose corresponding to the predicted anatomy is deformed onto the reference CT. This entire cycle is reiterated throughout the whole treatment time period, roughly 2-4 minutes, with the deformed predicted doses being accumulated on the reference 3D-CT. The total cumulative dose is then applied to the reference 3D-CT, facilitating the observation of differences compared to the planning CT.

Intra-fraction adaption requires prediction of target motion, and adjustment of radiation delivery in response, in real-time during the treatment. While our models compute motions quickly (inference time for each input image is ~27 msec and ~4 msec for GNN and CNN model, respectively), the inevitable response lag of the treatment system, due both to data processing following a motion input and the physical inertia of the delivery apparatus, means motion predictions must in practice be computed somewhat *ahead of time*. In principle, our approach could be extended to enable this using ideas from sequence-to-sequence learning, e.g. by incorporating RNN or LSTM components in the model to capture sequences of organ motions, given dynamic sequences of input kV images. A recent work has been done by our research group on generative modelling of cardiac MR image sequences [247] can provide further inspiration here.

Appendix A

Worst performing test cases (derived from SuPReMo) visualization for liver patients 2, 3 and 4 with GNN-based approach

Patient 2

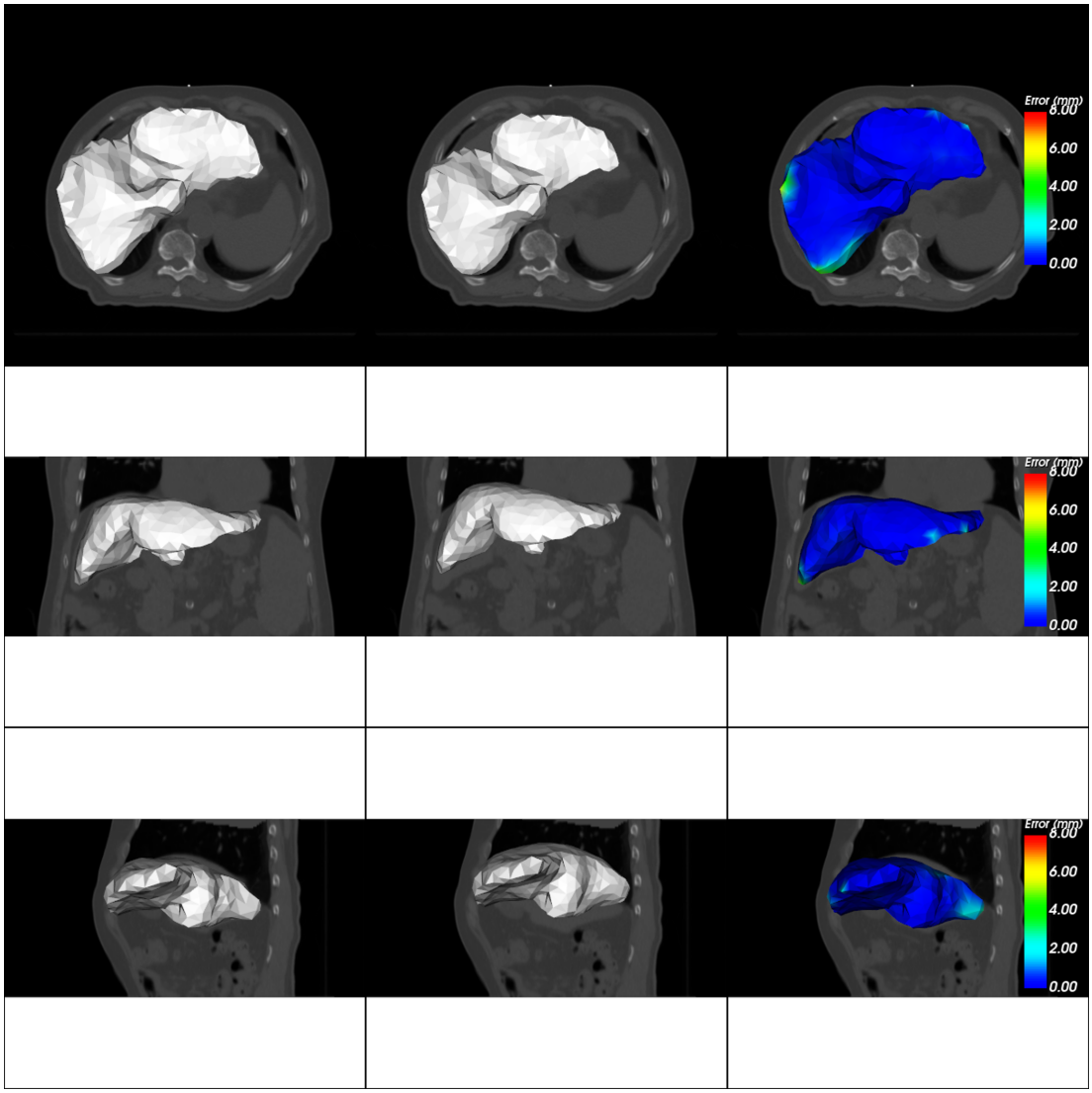


Figure A.1: Visualisations of ground-truth deformed (left column), template (middle column), and estimated deformed (right column) 3D liver shapes. Meshes are overlaid on the deformed 3D-CT volume. Rows 1-3 show, respectively, axial, coronal, and sagittal views. Results are shown for the *worst* performing test case for patient 2: image projection angle 23.195°, and deformation state producing highest errors. Contours in the right column indicate the spatial distribution of errors on the surface.

Patient 3

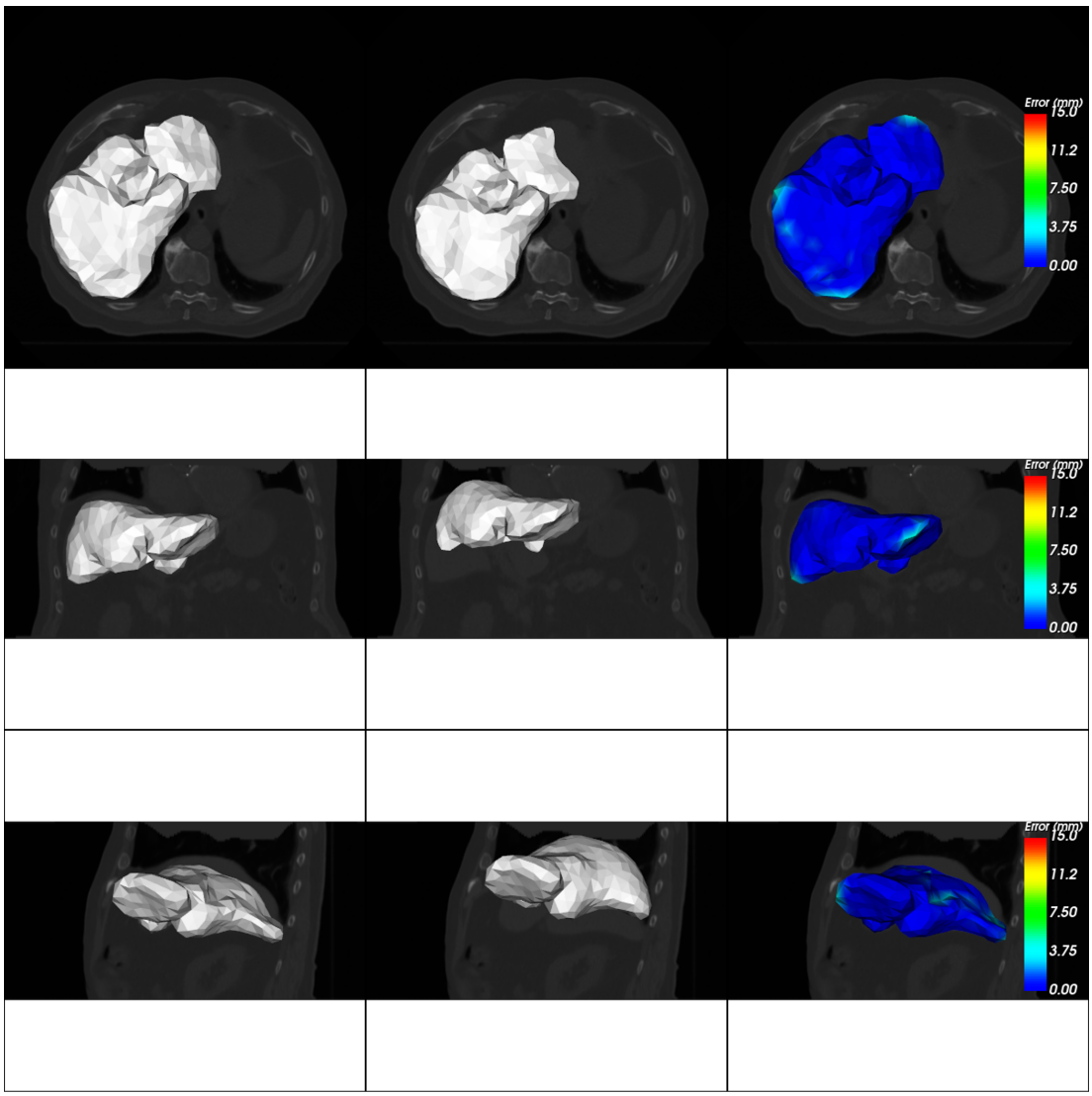


Figure A.2: Visualisations of ground-truth deformed (left column), template (middle column), and estimated deformed (right column) 3D liver shapes. Meshes are overlaid on the deformed 3D-CT volume. Rows 1-3 show, respectively, axial, coronal, and sagittal views. Results are shown for the *worst* performing test case for patient 3: image projection angle 202.009°, and deformation state producing highest errors. Contours in the right column indicate the spatial distribution of errors on the surface.

Patient 4

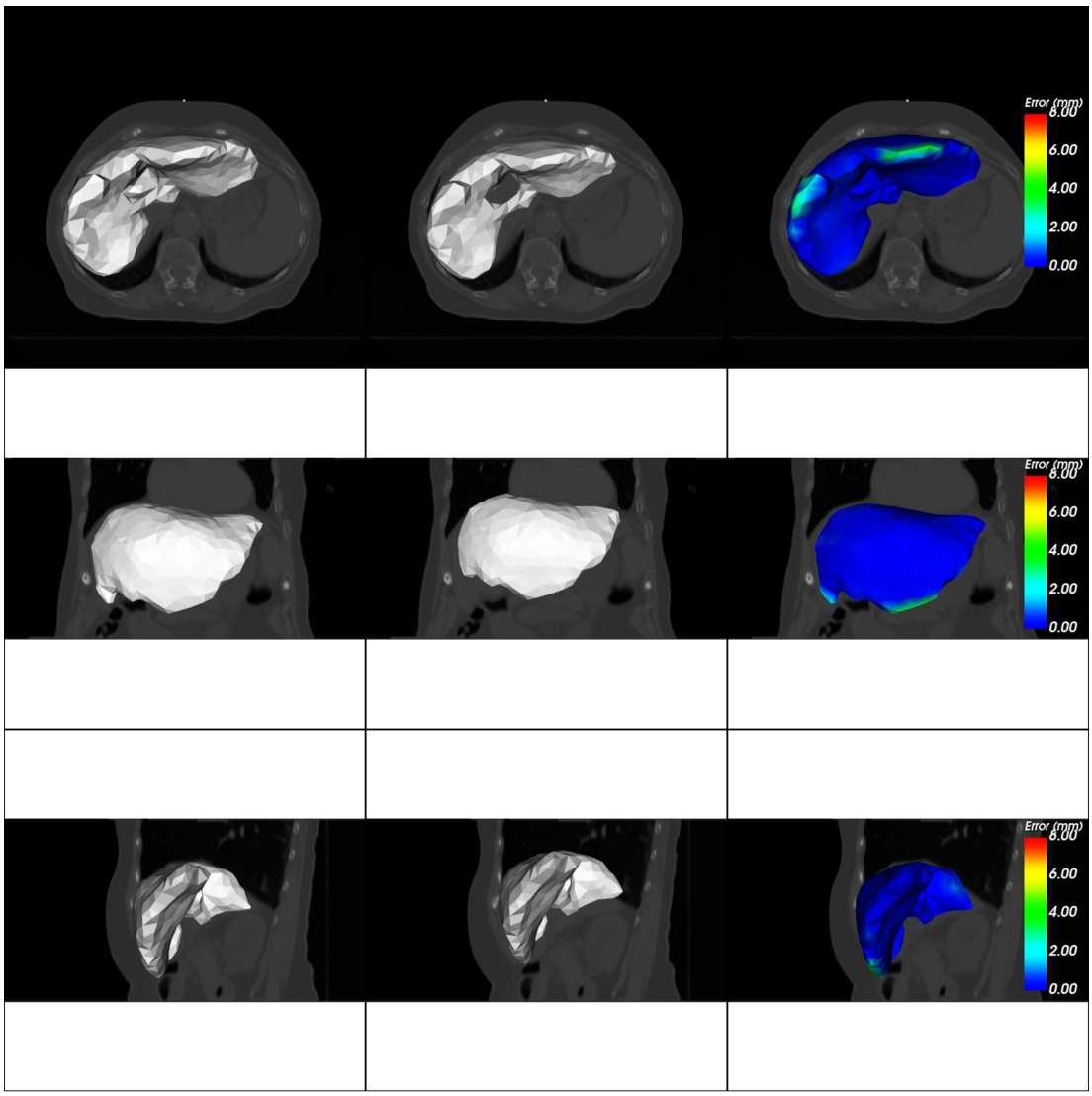


Figure A.3: Visualisations of ground-truth deformed (left column), template (middle column), and estimated deformed (right column) 3D liver shapes. Meshes are overlaid on the deformed 3D-CT volume. Rows 1-3 show, respectively, axial, coronal, and sagittal views. Results are shown for the *worst* performing test case for patient 4: image projection angle 256.387°, and deformation state producing highest errors. Contours in the right column indicate the spatial distribution of errors on the surface.

Appendix B

Worst performing test cases (derived from SuPReMo) visualization for liver patients 2, 3 and 4 with self-attention based CNN approach

Patient 2

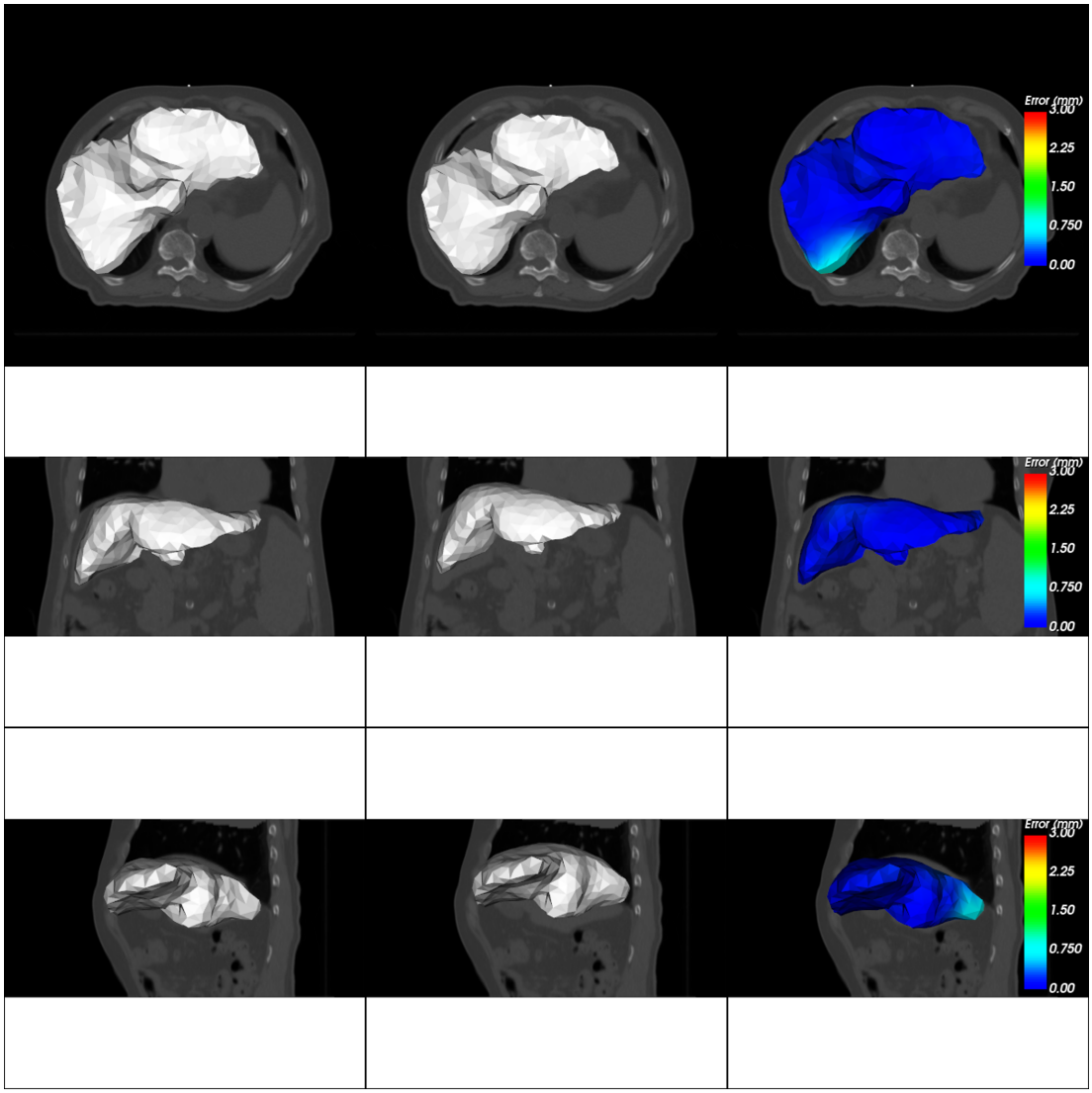
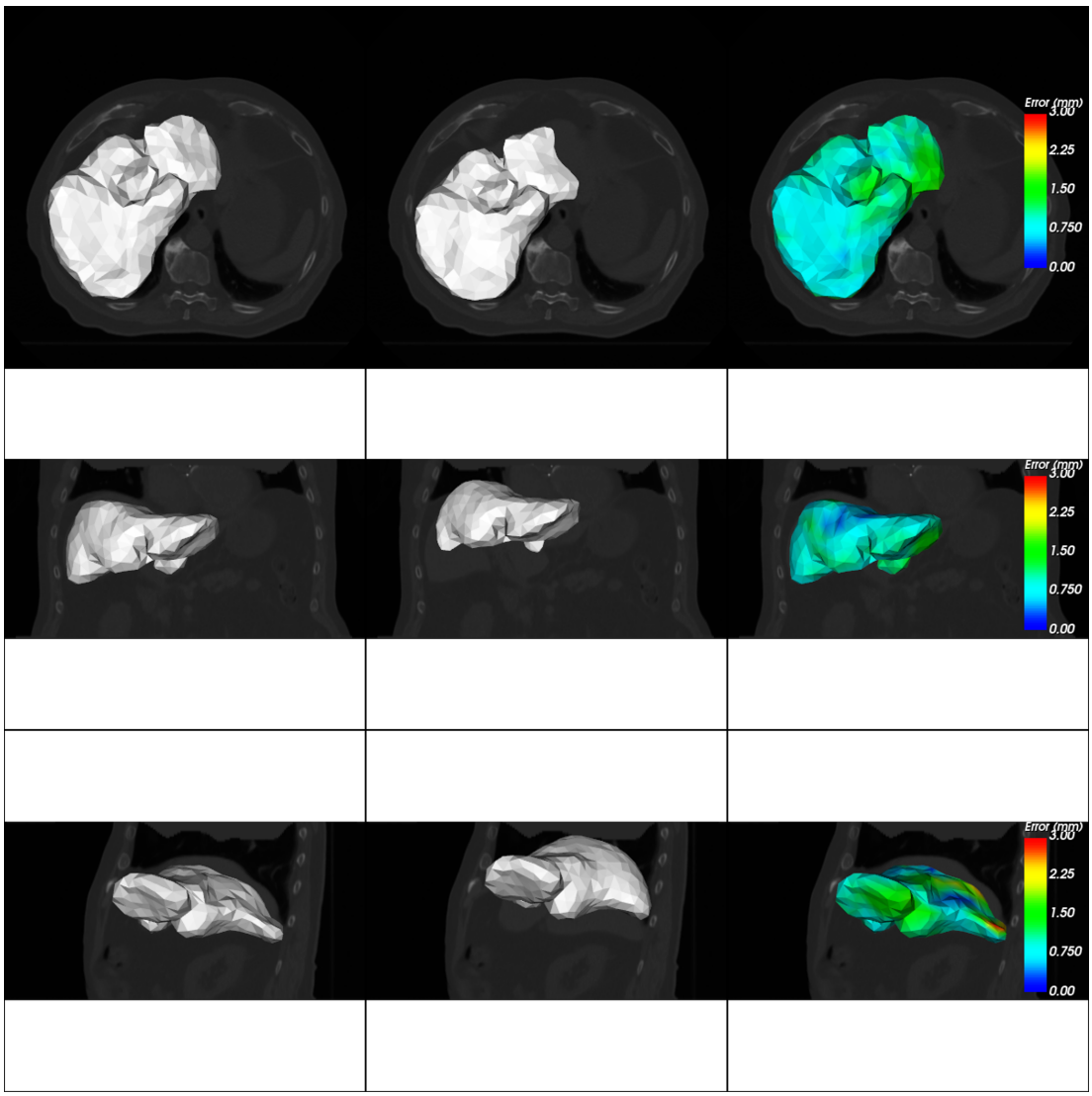


Figure B.1: Visualisations of ground-truth deformed (left column), template (middle column), and estimated deformed (right column) 3D liver shapes. Meshes are overlaid on the deformed 3D-CT volume. Rows 1-3 show, respectively, axial, coronal, and sagittal views. Results are shown for the *worst* performing test case for patient 2: image projection angle 214.586°, and deformation state producing highest errors. Contours in the right column indicate the spatial distribution of errors on the surface.

Patient 3



GT Mesh Alignment | Reference Mesh Alignment | Estimated Mesh at Projection Angle 358.246

Figure B.2: Visualisations of ground-truth deformed (left column), template (middle column), and estimated deformed (right column) 3D liver shapes. Meshes are overlaid on the deformed 3D-CT volume. Rows 1-3 show, respectively, axial, coronal, and sagittal views. Results are shown for the *worst* performing test case for patient 3: image projection angle 358.246°, and deformation state producing highest errors. Contours in the right column indicate the spatial distribution of errors on the surface.

Patient 4

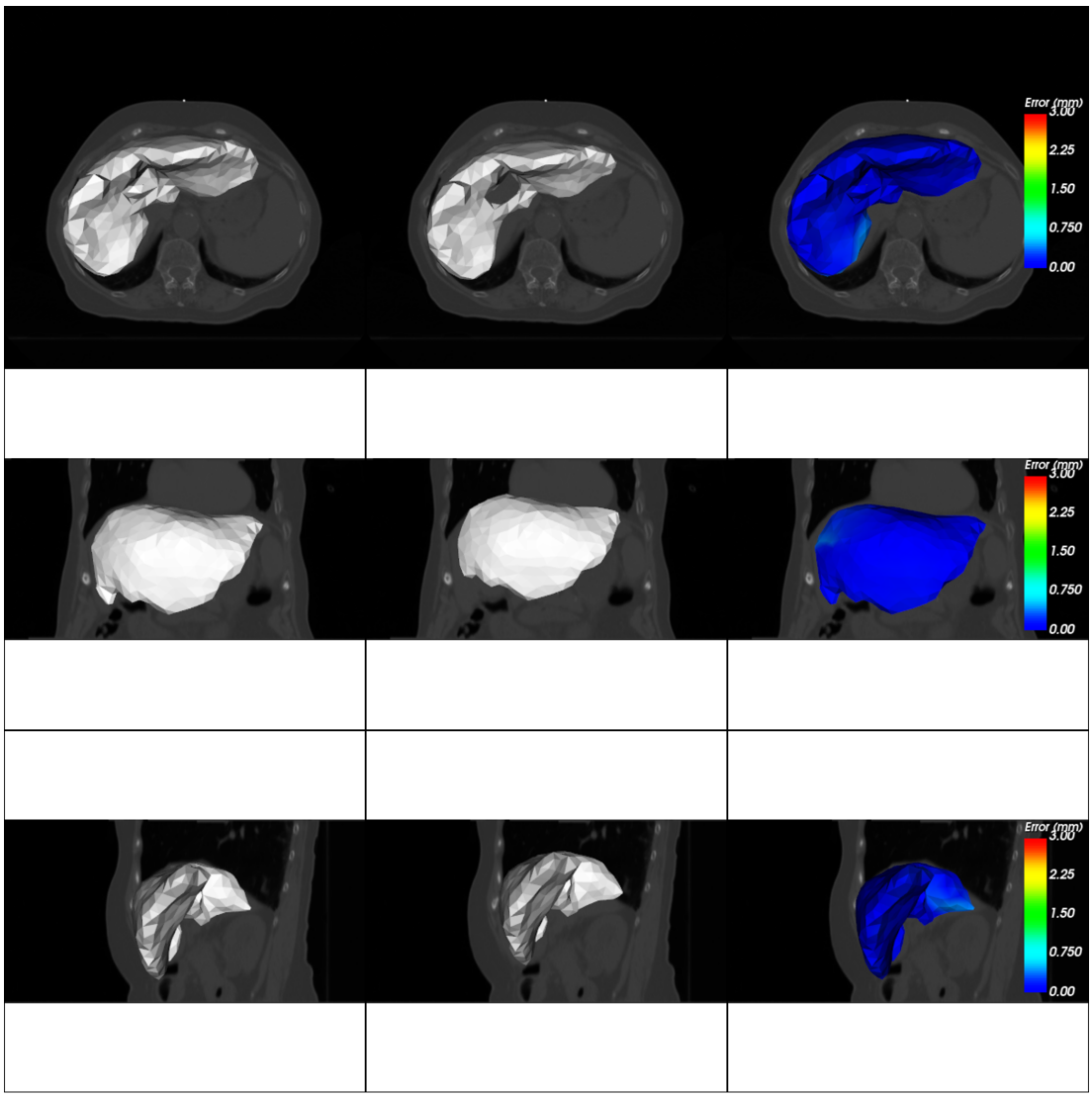


Figure B.3: Visualisations of ground-truth deformed (left column), template (middle column), and estimated deformed (right column) 3D liver shapes. Meshes are overlaid on the deformed 3D-CT volume. Rows 1-3 show, respectively, axial, coronal, and sagittal views. Results are shown for the *worst* performing test case for patient 4: image projection angle 106.955°, and deformation state producing highest errors. Contours in the right column indicate the spatial distribution of errors on the surface.

References

[1] Paul J Keall et al. “The management of respiratory motion in radiation oncology report of AAPM Task Group 76 a”. In: *Medical physics* 33.10 (2006), pp. 3874–3900.

[2] Rajamanickam Baskar et al. “Cancer and radiation therapy: current advances and future directions”. In: *International journal of medical sciences* 9.3 (2012), p. 193.

[3] You Zhang et al. “4D liver tumor localization using cone-beam projections and a biomechanical model”. In: *Radiotherapy and Oncology* 133 (2019), pp. 183–192.

[4] Alexander Chi, Nam Phong Nguyen, and Ritsuko Komaki. “The potential role of respiratory motion management and image guidance in the reduction of severe toxicities following stereotactic ablative radiation therapy for patients with centrally located early stage non-small cell lung cancer or lung metastases”. In: *Frontiers in oncology* 4 (2014), p. 151.

[5] M Gargett et al. “Clinical impact of removing respiratory motion during liver SABR”. In: *Radiation Oncology* 14.1 (2019), pp. 1–9.

[6] Hassan Abbas, Bryan Chang, and Zhe Jay Chen. “Motion management in gastrointestinal cancers”. In: *Journal of gastrointestinal oncology* 5.3 (2014), p. 223.

[7] Nikhil Bhagat et al. “Complications associated with the percutaneous insertion of fiducial markers in the thorax”. In: *Cardiovascular and interventional radiology* 33 (2010), pp. 1186–1191.

[8] Nicholas O Roman et al. “Interfractional positional variability of fiducial markers and primary tumors in locally advanced non-small-cell lung cancer during audiovisual biofeedback radiotherapy”. In: *International Journal of Radiation Oncology* Biology* Physics* 83.5 (2012), pp. 1566–1572.

[9] Jacob S Witt, Stephen A Rosenberg, and Michael F Bassetti. “MRI-guided adaptive radiotherapy for liver tumours: visualising the future”. In: *The Lancet Oncology* 21.2 (2020), e74–e82.

[10] Chiara Paganelli et al. “MRI-guidance for motion management in external beam radiotherapy: current status and future challenges”. In: *Physics in Medicine & Biology* 63.22 (2018), 22TR03.

[11] Paulina Krzyszczyk et al. “The growing role of precision and personalized medicine for cancer treatment”. In: *Technology* 6.03n04 (2018), pp. 79–100.

[12] Andrew Dhawan et al. “Tumour control probability in cancer stem cells hypothesis”. In: *PloS one* 9.5 (2014), e96093.

[13] Jenny Bertholet et al. “Real-time intrafraction motion monitoring in external beam radiotherapy”. In: *Physics in Medicine & Biology* 64.15 (2019), 15TR01.

[14] Walther Fledelius et al. “Tracking latency in image-based dynamic MLC tracking with direct image access”. In: *Acta Oncologica* 50.6 (2011), pp. 952–959.

[15] Hiroki Shirato et al. “Organ motion in image-guided radiotherapy: lessons from real-time tumor-tracking radiotherapy”. In: *International Journal of Clinical Oncology* 12.1 (2007), pp. 8–16.

[16] Mai Lykkegaard Schmidt et al. “Cardiac and respiration induced motion of mediastinal lymph node targets in lung cancer patients throughout the radiotherapy treatment course”. In: *Radiotherapy and Oncology* 121.1 (2016), pp. 52–58.

[17] Akira Imaizumi et al. “Transarterial fiducial marker implantation for CyberKnife radiotherapy to treat pancreatic cancer: an experience with 14 cases”. In: *Japanese journal of radiology* 39 (2021), pp. 84–92.

[18] Junichi Fukada et al. “Detection of esophageal fiducial marker displacement during radiation therapy with a 2-dimensional on-board imager: analysis of internal margin for esophageal cancer”. In: *International Journal of Radiation Oncology* Biology* Physics* 85.4 (2013), pp. 991–998.

[19] Silvia Carrara et al. “EUS-guided placement of fiducial markers for image-guided radiotherapy in gastrointestinal tumors: A critical appraisal”. In: *Endoscopic Ultrasound* 10.6 (2021), p. 414.

[20] Jean-Briac G Prévost et al. “Endovascular coils as lung tumour markers in real-time tumour tracking stereotactic radiotherapy: preliminary results”. In: *European radiology* 18.8 (2008), pp. 1569–1576.

[21] Hideki Hanazawa et al. “Clinical assessment of coiled fiducial markers as internal surrogates for hepatocellular carcinomas during gated stereotactic body radiotherapy with a real-time tumor-tracking system”. In: *Radiotherapy and Oncology* 123.1 (2017), pp. 43–48.

[22] Enrique Castellanos et al. “Low infection rate after transrectal implantation of Gold Anchor™ fiducial markers in prostate cancer patients after non-broad-spectrum antibiotic prophylaxis”. In: *Cureus* 10.10 (2018).

[23] Jonas Willmann et al. “Four-Dimensional Computed Tomography-Based Correlation of Respiratory Motion of Lung Tumors With Implanted Fiducials and an External Surrogate”. en. In: *Advances in Radiation Oncology* 7.3 (May 2022), p. 100885. ISSN: 24521094. DOI: 10.1016/j.adro.2021.100885. URL: https://linkinghub.elsevier.com/retrieve/pii/S2452109421002438 (visited on 11/25/2022).

[24] Hiroaki Akasaka et al. “Fiducial marker position affects target volume in stereotactic lung irradiation”. In: *Journal of Applied Clinical Medical Physics* 23.6 (2022), e13596.

[25] Daryl Lim Joon et al. “A clinical study comparing polymer and gold fiducials for prostate cancer radiotherapy”. In: *Frontiers in Oncology* 12 (2023). ISSN: 2234-943X. DOI: 10.3389/fonc.2022.1023288. URL: https://www.frontiersin.org/articles/10.3389/fonc.2022.1023288.

[26] Morgan Rose et al. “Bronchoscopic delivery of lipiodol as a fiducial marker in lung tumors before radiotherapy”. In: *Journal of Thoracic Oncology* 9.10 (2014), pp. 1579–1583.

[27] S. Sioshansi, L. Ding, and T.J. FitzGerald. “SBRT using residual Lipiodol as surrogate fiducial for image guidance in the treatment of recurrent or residual hepatocellular carcinoma”. In: *Practical Radiation Oncology* 3.2 (2013). DOI: 10.1016/j.prro.2013.01.061.

[28] Wei-Yang Lin et al. “Real-time automatic fiducial marker tracking in low contrast cine-MV images”. In: *Medical physics* 40.1 (2013), p. 011715.

[29] Margie A Hunt et al. “Simultaneous MV-kV imaging for intrafractional motion management during volumetric-modulated arc therapy delivery”. In: *Journal of applied clinical medical physics* 17.2 (2016), pp. 473–486.

[30] Xiaoli Tang, Greg C Sharp, and Steve B Jiang. “Fluoroscopic tracking of multiple implanted fiducial markers using multiple object tracking”. In: *Physics in Medicine & Biology* 52.14 (2007), p. 4081.

[31] W Mao, RD Wiersma, and L Xing. “Fast internal marker tracking algorithm for onboard MV and kV imaging systems”. In: *Medical physics* 35.5 (2008), pp. 1942–1949.

[32] Thomas E Marchant, Andrzej Skalski, and BJ Matuszewski. “Automatic tracking of implanted fiducial markers in cone beam CT projection images”. In: *Medical physics* 39.3 (2012), pp. 1322–1334.

[33] Walther Fledelius et al. “Real-time segmentation of multiple implanted cylindrical liver markers in kilovoltage and megavoltage x-ray images”. In: *Physics in Medicine & Biology* 59.11 (2014), p. 2787.

[34] Rajesh Regmi et al. “Automatic tracking of arbitrarily shaped implanted markers in kilovoltage projection images: a feasibility study”. In: *Medical physics* 41.7 (2014), p. 071906.

[35] Jenny Bertholet et al. “Fully automatic segmentation of arbitrarily shaped fiducial markers in cone-beam CT projections”. In: *Physics in Medicine & Biology* 62.4 (2017), p. 1327.

[36] Hanlin Wan et al. “Automated patient setup and gating using cone beam computed tomography projections”. In: *Physics in Medicine & Biology* 61.6 (2016), p. 2552.

[37] Nishita Kothary et al. “Safety and efficacy of percutaneous fiducial marker implantation for image-guided radiation therapy”. In: *Journal of vascular and interventional radiology* 20.2 (2009), pp. 235–239.

[38] Anna Kirilova et al. “Three-dimensional motion of liver tumors using cine-magnetic resonance imaging”. In: *International Journal of Radiation Oncology* Biology* Physics* 71.4 (2008), pp. 1189–1195.

[39] John R Adler Jr et al. “The Cyberknife: a frameless robotic system for radiosurgery”. In: *Stereotactic and functional neurosurgery* 69.1-4 (1997), pp. 124–128.

[40] Martin J Murphy et al. “Image-guided radiosurgery for the spine and pancreas”. In: *Computer Aided Surgery* 5.4 (2000), pp. 278–288.

[41] Jay L Friedland et al. “Stereotactic body radiotherapy: an emerging treatment approach for localized prostate cancer”. In: *Technology in cancer research & treatment* 8.5 (2009), pp. 387–392.

[42] Christopher R King et al. “Long-term outcomes from a prospective trial of stereotactic body radiotherapy for low-risk prostate cancer”. In: *International Journal of Radiation Oncology* Biology* Physics* 82.2 (2012), pp. 877–882.

[43] Marcello Serra et al. “SBRT for localized prostate cancer: CyberKnife vs. VMAT-FFF, a dosimetric study”. In: *Life* 12.5 (2022), p. 711.

[44] Ahmet Murat Şenişik et al. “A dosimetric comparison for SBRT plans of localized prostate cancer between Cyberknife and Varian Truebeam STX device”. In: *Applied Radiation and Isotopes* 192 (2023), p. 110617.

[45] Makoto Ito et al. “Stereotactic body radiation therapy for prostate cancer: a study comparing 3-year genitourinary toxicity between CyberKnife and volumetric-

modulated arc therapy by propensity score analysis". In: *Radiation Oncology* 18.1 (2023), pp. 1–10.

[46] Hiroki Shirato and Shinichi Shimizu. "Real-time tumour-tracking radiotherapy." In: *The Lancet* 353.9161 (1999), pp. 1331–1332.

[47] Rumiko Kinoshita et al. "Three-dimensional intrafractional motion of breast during tangential breast irradiation monitored with high-sampling frequency using a real-time tumor-tracking radiotherapy system". In: *International Journal of Radiation Oncology* Biology* Physics* 70.3 (2008), pp. 931–934.

[48] Takehiro Shiinoki et al. "Verification of respiratory-gated radiotherapy with new real-time tumour-tracking radiotherapy system using cine EPID images and a log file". In: *Physics in Medicine & Biology* 62.4 (2017), p. 1585.

[49] Yuichiro Kamino et al. "Development of a four-dimensional image-guided radiotherapy system with a gimbaled X-ray head". In: *International Journal of Radiation Oncology* Biology* Physics* 66.1 (2006), pp. 271–278.

[50] Shinichiro Mori et al. "Carbon-ion pencil beam scanning treatment with gated markerless tumor tracking: an analysis of positional accuracy". In: *International Journal of Radiation Oncology* Biology* Physics* 95.1 (2016), pp. 258–266.

[51] Marco Serpa and Christoph Bert. "Dense feature-based motion estimation in MV fluoroscopy during dynamic tumor tracking treatment: preliminary study on reduced aperture and partial occlusion handling". In: *Physics in Medicine & Biology* 65.24 (2020), p. 245039.

[52] D. Roa et al. "Monte Carlo simulations and phantom validation of low-dose radiotherapy to the lungs using an interventional radiology C-arm fluoroscope". In: *Physica Medica* 94 (2022), pp. 24–34. ISSN: 1120-1797. DOI: https://

doi . org / 10 . 1016 / j . ejmp . 2021 . 12 . 014. URL: https : / / www . sciencedirect.com/science/article/pii/S1120179721003677.

[53] D Ferguson et al. "Automated MV markerless tumor tracking for VMAT". In: *Physics in Medicine & Biology* 65.12 (2020), p. 125011.

[54] Kimmie de Bruin et al. "Markerless Real-Time 3-Dimensional kV Tracking of Lung Tumors During Free Breathing Stereotactic Radiation Therapy". In: *Advances in Radiation Oncology* 6.4 (2021), p. 100705. ISSN: 2452-1094. DOI: https://doi.org/10.1016/j.adro.2021.100705. URL: https://www.sciencedirect.com/science/article/pii/S2452109421000634.

[55] Marco Mueller et al. "MArkerless image Guidance using Intrafraction Kilovoltage x-ray imaging (MAGIK): study protocol for a phase I interventional study for lung cancer radiotherapy". In: *BMJ Open* 12.1 (2022). ISSN: 2044-6055. DOI: 10.1136/bmjopen-2021-057135. eprint: https://bmjopen.bmj.com/content/12/1/e057135.full.pdf. URL: https://bmjopen.bmj.com/content/12/1/e057135.

[56] Bastian L Lindl et al. "TOPOS: a new topometric patient positioning and tracking system for radiation therapy based on structured white light". In: *Medical physics* 40.4 (2013), p. 042701.

[57] Simon Placht et al. "Fast time-of-flight camera based surface registration for radiotherapy patient positioning". In: *Medical physics* 39.1 (2012), pp. 4–17.

[58] Anders Brahme, Peter Nyman, and Björn Skatt. "4D laser camera for accurate patient positioning, collision avoidance, image fusion and adaptive approaches during diagnostic and therapeutic procedures". In: *Medical Physics* 35.5 (2008), pp. 1670–1681.

[59] Christoph Bert et al. “A phantom evaluation of a stereo-vision surface imaging system for radiotherapy patient setup”. In: *Medical physics* 32.9 (2005), pp. 2753–2762.

[60] Jian-Yue Jin et al. “Use of the BrainLAB ExacTrac X-Ray 6D system in image-guided radiotherapy”. In: *Medical Dosimetry* 33.2 (2008), pp. 124–134.

[61] Twyla Willoughby et al. “Quality assurance for nonradiographic radiotherapy localization and positioning systems: report of Task Group 147”. In: *Medical physics* 39.4 (2012), pp. 1728–1747.

[62] Ben Perrett et al. “A framework for exactrac dynamic commissioning for stereotactic radiosurgery and stereotactic ablative radiotherapy”. In: *Journal of Medical Physics* 47.4 (2022), p. 398.

[63] Ruijiang Li et al. “Evaluation of the geometric accuracy of surrogate-based gated VMAT using intrafraction kilovoltage x-ray images”. In: *Medical physics* 39.5 (2012), pp. 2686–2693.

[64] Jonas Scherman Rydhög et al. “Target position uncertainty during visually guided deep-inspiration breath-hold radiotherapy in locally advanced lung cancer”. In: *Radiotherapy and Oncology* 123.1 (2017), pp. 78–84.

[65] Aurora Fassi et al. “Target position reproducibility in left-breast irradiation with deep inspiration breath-hold using multiple optical surface control points”. In: *Journal of applied clinical medical physics* 19.4 (2018), pp. 35–43.

[66] Yevgeniy Vinogradskiy et al. “The clinical and dosimetric impact of real-time target tracking in pancreatic SBRT”. In: *International Journal of Radiation Oncology* Biology* Physics* 103.1 (2019), pp. 268–275.

[67] Jason K Molitoris et al. “Advances in the use of motion management and image guidance in radiation therapy treatment for lung cancer”. In: *Journal of thoracic disease* 10.Suppl 21 (2018), S2437.

[68] Guang Li et al. “Motion monitoring for cranial frameless stereotactic radiosurgery using video-based three-dimensional optical surface imaging”. In: *Medical physics* 38.7 (2011), pp. 3981–3994.

[69] Hubert Pan et al. “Frameless, real-time, surface imaging-guided radiosurgery: clinical outcomes for brain metastases”. In: *Neurosurgery* 71.4 (2012), pp. 844–852.

[70] Jeremy DP Hoisak and Todd Pawlicki. “The role of optical surface imaging systems in radiation therapy”. In: *Seminars in radiation oncology*. Vol. 28. Elsevier. 2018, pp. 185–193.

[71] Xiaoli Tang et al. “Clinical experience with 3-dimensional surface matching-based deep inspiration breath hold for left-sided breast cancer radiation therapy”. In: *Practical radiation oncology* 4.3 (2014), e151–e158.

[72] Zhao Ma et al. “Optical surface management system for patient positioning in interfractional breast cancer radiotherapy”. In: *BioMed research international* 2018 (2018).

[73] Konstantin Christoph Koban et al. “Three-dimensional surface imaging in breast cancer: a new tool for clinical studies?” In: *Radiation Oncology* 15.1 (2020), pp. 1–8.

[74] Albert J Chang et al. “Video surface image guidance for external beam partial breast irradiation”. In: *Practical Radiation Oncology* 2.2 (2012), pp. 97–105.

[75] Amish P Shah et al. “Clinical evaluation of interfractional variations for whole breast radiotherapy using 3-dimensional surface imaging”. In: *Practical radiation oncology* 3.1 (2013), pp. 16–25.

[76] David B Wiant et al. “Surface imaging-based analysis of intrafraction motion for breast radiotherapy patients”. In: *Journal of applied clinical medical physics* 15.6 (2014), pp. 147–159.

[77] Marko Laaksomaa et al. “AlignRT® and Catalyst™ in whole-breast radiotherapy with DIBH: Is IGRT still needed?” In: *Journal of applied clinical medical physics* 20.3 (2019), pp. 97–104.

[78] Veronica Sorgato et al. “Benchmarking the AlignRT surface deformation module for the early detection and quantification of oedema in breast cancer radiotherapy”. In: *Technical Innovations & Patient Support in Radiation Oncology* 21 (2022), pp. 16–22.

[79] Guang Li et al. “A uniform and versatile surface-guided radiotherapy procedure and workflow for high-quality breast deep-inspiration breath-hold treatment in a multi-center institution”. In: *Journal of applied clinical medical physics* 23.3 (2022), e13511.

[80] Jeremy DP Hoisak et al. “Correlation of lung tumor motion with external surrogate indicators of respiration”. In: *International Journal of Radiation Oncology* Biology* Physics* 60.4 (2004), pp. 1298–1306.

[81] John W Wong et al. “The use of active breathing control (ABC) to reduce margin for breathing motion”. In: *International Journal of Radiation Oncology* Biology* Physics* 44.4 (1999), pp. 911–919.

[82] Helen A McNair et al. “Feasibility of the use of the Active Breathing Co ordinator™(ABC) in patients receiving radical radiotherapy for non-small cell

lung cancer (NSCLC)". In: *Radiotherapy and Oncology* 93.3 (2009), pp. 424–429.

[83] Vincent M Remouchamps et al. "Three-dimensional evaluation of intra-and interfraction immobilization of lung and chest wall using active breathing control: a reproducibility study with breast cancer patients". In: *International Journal of Radiation Oncology* Biology* Physics* 57.4 (2003), pp. 968–978.

[84] Anders N Pedersen et al. "Breathing adapted radiotherapy of breast cancer: reduction of cardiac and pulmonary doses using voluntary inspiration breath-hold". In: *Radiotherapy and oncology* 72.1 (2004), pp. 53–60.

[85] Stine S Korreman et al. "Breathing adapted radiotherapy for breast cancer: comparison of free breathing gating with the breath-hold technique". In: *Radiotherapy and oncology* 76.3 (2005), pp. 311–318.

[86] Laura A Dawson et al. "Accuracy of daily image guidance for hypofractionated liver radiotherapy with active breathing control". In: *International Journal of Radiation Oncology* Biology* Physics* 62.4 (2005), pp. 1247–1252.

[87] Cynthia Eccles et al. "Reproducibility of liver position using active breathing coordinator for liver cancer radiotherapy". In: *International Journal of Radiation Oncology* Biology* Physics* 64.3 (2006), pp. 751–759.

[88] Marianne Camille Aznar et al. "ESTRO-ACROP guideline: Recommendations on implementation of breath-hold techniques in radiotherapy". In: *Radiotherapy and Oncology* 185 (2023), p. 109734.

[89] Jeremy DP Hoisak et al. "Correlation of lung tumor motion with external surrogate indicators of respiration". In: *International Journal of Radiation Oncology* Biology* Physics* 60.4 (2004), pp. 1298–1306.

[90] Cihat Ozhasoglu et al. “Synchrony–cyberknife respiratory compensation technology”. In: *Medical Dosimetry* 33.2 (2008), pp. 117–123.

[91] Mischa Hoogeman et al. “Clinical accuracy of the respiratory tumor tracking system of the cyberknife: assessment by analysis of log files”. In: *International Journal of Radiation Oncology* Biology* Physics* 74.1 (2009), pp. 297–303.

[92] Jean-Emmanuel Bibault et al. “Image-guided robotic stereotactic radiation therapy with fiducial-free tumor tracking for lung cancer”. In: *Radiation Oncology* 7.1 (2012), p. 102.

[93] William S Ferris et al. “Evaluation of radixact motion synchrony for 3D respiratory motion: Modeling accuracy and dosimetric fidelity”. In: *Journal of Applied Clinical Medical Physics* 21.9 (2020), pp. 96–106.

[94] Mei Yan Tse et al. “Dosimetric impact of phase shifts on Radixact Synchrony tracking system with patient-specific breathing patterns”. In: *Journal of Applied Clinical Medical Physics* 23.6 (2022), e13600.

[95] Twyla R Willoughby et al. “Evaluation of an infrared camera and X-ray system using implanted fiducials in patients with lung tumors for gated radiation therapy”. In: *International Journal of Radiation Oncology* Biology* Physics* 66.2 (2006), pp. 568–575.

[96] Jian-Yue Jin et al. “Use of the BrainLAB ExacTrac X-Ray 6D system in image-guided radiotherapy”. In: *Medical Dosimetry* 33.2 (2008), pp. 124–134.

[97] Simon K Goodall and Peter L Rampant. “Initial end-to-end testing of the ExacTrac dynamic deep inspiration breath hold workflow using a breath hold breast phantom”. In: *Physical and Engineering Sciences in Medicine* (2023), pp. 1–9.

[98] Tom Depuydt et al. “Treating patients with real-time tumor tracking using the Vero gimbaled linac system: implementation and first review”. In: *Radiotherapy and Oncology* 112.3 (2014), pp. 343–351.

[99] Mami Akimoto et al. “Predictive uncertainty in infrared marker-based dynamic tumor tracking with Vero4DRT a”. In: *Medical physics* 40.9 (2013), p. 091705.

[100] R Orecchia et al. “VERO® radiotherapy for low burden cancer: 789 patients with 957 lesions”. In: *ecancermedicalscience* 10 (2016).

[101] Ross I Berbeco et al. “Residual motion of lung tumours in gated radiotherapy with external respiratory surrogates”. In: *Physics in Medicine & Biology* 50.16 (2005), p. 3655.

[102] Maggie Margaret Anderson. “Investigating the robustness of the Anzai respiratory gating system”. In: (2013).

[103] Jenny Bertholet et al. “Automatic online and real-time tumour motion monitoring during stereotactic liver treatments on a conventional linac by combined optical and sparse monoscopic imaging with kilovoltage x-rays (COSMIK)”. In: *Physics in Medicine & Biology* 63.5 (2018), p. 055012.

[104] Per Rugaard Poulsen, Byungchul Cho, and Paul J Keall. “A method to estimate mean position, motion magnitude, motion correlation, and trajectory of a tumor from cone-beam CT projections for image-guided radiotherapy”. In: *International Journal of Radiation Oncology* Biology* Physics* 72.5 (2008), pp. 1587–1596.

[105] D Ruan et al. “Inference of hysteretic respiratory tumor motion from external surrogates: a state augmentation approach”. In: *Physics in Medicine & Biology* 53.11 (2008), p. 2923.

[106] Thomas Ravkilde et al. “First online real-time evaluation of motion-induced 4D dose errors during radiotherapy delivery”. In: *Medical physics* 45.8 (2018), pp. 3893–3903.

[107] S Skouboe et al. “OC-0543 First clinical real-time motion-including tumor dose reconstruction during radiotherapy delivery”. In: *Radiotherapy and Oncology* 133 (2019), S286–S287.

[108] Ahmad Amoush et al. “Single-isocenter hybrid IMRT plans versus two-isocenter conventional plans and impact of intrafraction motion for the treatment of breast cancer with supraclavicular lymph nodes involvement”. In: *Journal of Applied Clinical Medical Physics* 16.4 (2015), pp. 31–39.

[109] Xiaoying Liang et al. “Using robust optimization for skin flashing in intensity modulated radiation therapy for breast cancer treatment: A feasibility study”. In: *Practical Radiation Oncology* 10.1 (2020), pp. 59–69.

[110] Yuqing Xiong et al. “Assessment of intrafractional prostate motion and its dosimetric impact in MRI-guided online adaptive radiotherapy with gating”. In: *Strahlentherapie und Onkologie* (2022), pp. 1–10.

[111] Saber Nankali et al. “Intrafraction tumor motion monitoring and dose reconstruction for liver pencil beam scanning proton therapy”. In: *Frontiers in oncology* 13 (2023).

[112] James M Balter et al. “Accuracy of a wireless localization system for radiotherapy”. In: *International Journal of Radiation Oncology* Biology* Physics* 61.3 (2005), pp. 933–937.

[113] Twyla R Willoughby et al. “Target localization and real-time tracking using the Calypso 4D localization system in patients with localized prostate cancer”. In:

International Journal of Radiation Oncology Biology* Physics* 65.2 (2006), pp. 528–534.

[114] Patrick Kupelian et al. "Multi-institutional clinical experience with the Calypso System in localization and continuous, real-time monitoring of the prostate gland during external radiotherapy". In: *International Journal of Radiation Oncology* Biology* Physics* 67.4 (2007), pp. 1088–1098.

[115] Eric T Shinohara et al. "Feasibility of electromagnetic transponder use to monitor inter-and intrafractional motion in locally advanced pancreatic cancer patients". In: *International Journal of Radiation Oncology* Biology* Physics* 83.2 (2012), pp. 566–573.

[116] Amish P Shah et al. "Real-time tumor tracking in the lung using an electromagnetic tracking system". In: *International Journal of Radiation Oncology* Biology* Physics* 86.3 (2013), pp. 477–483.

[117] Daniela Schmitt et al. "Motion monitoring during a course of lung radiotherapy with anchored electromagnetic transponders". In: *Strahlentherapie und Onkologie* 193.10 (2017), pp. 840–847.

[118] A Vanhanen, P Poulsen, and M Kapanen. "Dosimetric effect of intrafraction motion and different localization strategies in prostate SBRT". In: *Physica Medica* 75 (2020), pp. 58–68.

[119] Dante PI Capaldi et al. "A robotically assisted 3D printed quality assurance lung phantom for Calypso". In: *Physics in Medicine & Biology* 66.7 (2021), p. 074005.

[120] Ye Zhang et al. "Deformable motion reconstruction for scanned proton beam therapy using on-line x-ray imaging". In: *Physics in Medicine & Biology* 58.24 (2013), p. 8621.

[121] AK Richardson and P Jacobs. “Intrafraction monitoring of prostate motion during radiotherapy using the Clarity® Autoscan Transperineal Ultrasound (TPUS) system”. In: *Radiography* 23.4 (2017), pp. 310–313.

[122] Svenja Ipsen et al. “Online 4D ultrasound guidance for real-time motion compensation by MLC tracking”. In: *Medical physics* 43.10 (2016), pp. 5695–5704.

[123] Minglun Li et al. “Comparison of prostate positioning guided by three-dimensional transperineal ultrasound and cone beam CT”. In: *Strahlentherapie und Onkologie* 193.3 (2017), p. 221.

[124] Dwi Seno Kuncoro Sihono et al. “Determination of intrafraction prostate motion during external beam radiation therapy with a transperineal 4-dimensional ultrasound real-time tracking system”. In: *International Journal of Radiation Oncology* Biology* Physics* 101.1 (2018), pp. 136–143.

[125] Svenja Ipsen et al. “Simultaneous acquisition of 4D ultrasound and wireless electromagnetic tracking for in-vivo accuracy validation”. In: *Current Directions in Biomedical Engineering* 3.2 (2017), pp. 75–78.

[126] Dwi Seno et al. “A 4D ultrasound real-time tracking system for external beam radiotherapy of upper abdominal lesions under breath-hold”. In: *Strahlentherapie und Onkologie* 193.3 (2017), p. 213.

[127] Lena Vogel et al. “Intra-breath-hold residual motion of image-guided DIBH liver-SBRT: an estimation by ultrasound-based monitoring correlated with diaphragm position in CBCT”. In: *Radiotherapy and Oncology* 129.3 (2018), pp. 441–448.

[128] Tianlong Ji et al. "A phantom-based analysis for tracking intra-fraction pancreatic tumor motion by ultrasound imaging during radiation therapy". In: *Frontiers in Oncology* 12 (2022), p. 996537.

[129] Marie Fargier-Voiron et al. "Evaluation of a new transperineal ultrasound probe for inter-fraction image-guidance for definitive and post-operative prostate cancer radiotherapy". In: *Physica Medica* 32.3 (2016), pp. 499–505.

[130] Minglun Li et al. "Prefraction displacement and intrafraction drift of the prostate due to perineal ultrasound probe pressure". In: *Strahlentherapie und onkologie* 193.6 (2017), pp. 459–465.

[131] Hasan Tutkun Şen et al. "System integration and in vivo testing of a robot for ultrasound guidance and monitoring during radiotherapy". In: *IEEE Transactions on Biomedical Engineering* 64.7 (2016), pp. 1608–1618.

[132] Matt A Bernstein, Kevin F King, and Xiaohong Joe Zhou. *Handbook of MRI pulse sequences*. Elsevier, 2004.

[133] Brian Hargreaves. "Rapid gradient-echo imaging". In: *Journal of Magnetic Resonance Imaging* 36.6 (2012), pp. 1300–1313.

[134] Anagha Deshmane et al. "Parallel MR imaging". In: *Journal of Magnetic Resonance Imaging* 36.1 (2012), pp. 55–72.

[135] Bjorn Stemkens et al. "Image-driven, model-based 3D abdominal motion estimation for MR-guided radiotherapy". In: *Physics in Medicine & Biology* 61.14 (2016), p. 5335.

[136] EH Tran et al. "OC-0411: investigation of MRI-derived surrogate signals for modelling respiratory motion on an MRI-Linac". In: *Radiotherapy and Oncology* 127 (2018), S211–S212.

[137] Alexandra E Bourque et al. “A particle filter based autocontouring algorithm for lung tumor tracking using dynamic magnetic resonance imaging”. In: *Medical physics* 43.9 (2016), pp. 5161–5169.

[138] Eugene Yip et al. “Evaluating performance of a user-trained MR lung tumor autocontouring algorithm in the context of intra-and interobserver variations”. In: *Medical physics* 45.1 (2018), pp. 307–313.

[139] Yan Wang et al. “Deep learning based fully automatic segmentation of the left ventricular endocardium and epicardium from cardiac cine MRI”. In: *Quantitative Imaging in Medicine and Surgery* 11.4 (2021), p. 1600.

[140] Shunjie Dong et al. “DeU-Net 2.0: Enhanced deformable U-Net for 3D cardiac cine MRI segmentation”. In: *Medical Image Analysis* 78 (2022), p. 102389.

[141] Martin J Menten et al. “The impact of 2D cine MR imaging parameters on automated tumor and organ localization for MR-guided real-time adaptive radiotherapy”. In: *Physics in Medicine & Biology* 63.23 (2018), p. 235005.

[142] Xingyu Nie and Guang Li. “Real-Time 2D MR Cine From Beam Eye’s View With Tumor-Volume Projection to Ensure Beam-to-Tumor Conformality for MR-Guided Radiotherapy of Lung Cancer”. In: *Frontiers in Oncology* 12 (2022), p. 898771.

[143] Olga L Green et al. “First clinical implementation of real-time, real anatomy tracking and radiation beam control”. In: *Medical physics* 45.8 (2018), pp. 3728–3740.

[144] Lauren Henke et al. “Phase I trial of stereotactic MR-guided online adaptive radiation therapy (SMART) for the treatment of oligometastatic or unresectable primary malignancies of the abdomen”. In: *Radiotherapy and Oncology* 126.3 (2018), pp. 519–526.

[145] Shyama Tetar et al. “Patient-reported outcome measurements on the tolerance of magnetic resonance imaging-guided radiation therapy”. In: *Cureus* 10.2 (2018).

[146] Benjamin W Fischer-Valuck et al. “Two-and-a-half-year clinical experience with the world’s first magnetic resonance image guided radiation therapy system”. In: *Advances in radiation oncology* 2.3 (2017), pp. 485–493.

[147] Sahaja Acharya et al. “Online magnetic resonance image guided adaptive radiation therapy: first clinical applications”. In: *International Journal of Radiation Oncology* Biology* Physics* 94.2 (2016), pp. 394–403.

[148] Evan Liang et al. “Application of Continuous Positive Airway Pressure for Thoracic Respiratory Motion Management: An Assessment in a Magnetic Resonance Imaging–Guided Radiation Therapy Environment”. In: *Advances in Radiation Oncology* 7.3 (2022), p. 100889.

[149] Prescilla Uijtewaal et al. “Dosimetric evaluation of MRI-guided multi-leaf collimator tracking and trailing for lung stereotactic body radiation therapy”. In: *Medical Physics* 48.4 (2021), pp. 1520–1532.

[150] Ergys Subashi et al. “View-sharing for 4D magnetic resonance imaging with randomized projection-encoding enables improvements of respiratory motion imaging for treatment planning in abdominothoracic radiotherapy”. In: *Physics and Imaging in Radiation Oncology* (2023).

[151] Agnès Tallet et al. “Is MRI-Linac helpful in SABR treatments for liver cancer?” In: *Frontiers in Oncology* 13 (2023). ISSN: 2234-943X. DOI: 10.3389/fonc.2023.1130490. URL: https://www.frontiersin.org/articles/10.3389/fonc.2023.1130490.

[152] Peter Fischer et al. “Unsupervised learning for robust respiratory signal estimation from X-ray fluoroscopy”. In: *IEEE transactions on medical imaging* 36.4 (2016), pp. 865–877.

[153] A Balasubramanian et al. “Predictive modeling of respiratory tumor motion for real-time prediction of baseline shifts”. In: *Physics in Medicine & Biology* 62.5 (2017), p. 1791.

[154] Ryusuke Hirai et al. “Real-time tumor tracking using fluoroscopic imaging with deep neural network analysis”. In: *Physica Medica* 59 (2019), pp. 22–29.

[155] Geert Litjens et al. “A survey on deep learning in medical image analysis”. In: *Medical image analysis* 42 (2017), pp. 60–88.

[156] Jiang Kai, Fumitake Fujii, and Takehiro Shiinoki. “Prediction of Lung Tumor Motion Based on Recurrent Neural Network”. In: *2018 IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE. 2018, pp. 1093–1099.

[157] Seonyeong Park et al. “Intra-and inter-fractional variation prediction of lung tumors using fuzzy deep learning”. In: *IEEE journal of translational engineering in health and medicine* 4 (2016), pp. 1–12.

[158] Ran Wang et al. “A feasibility of respiration prediction based on deep Bi-LSTM for real-time tumor tracking”. In: *IEEE Access* 6 (2018), pp. 51262–51268.

[159] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[160] Elisabeth Steiner et al. “Both four-dimensional computed tomography and four-dimensional cone beam computed tomography under-predict lung tar-

get motion during radiotherapy". In: *Radiotherapy and Oncology* 135 (2019), pp. 65–73.

[161] Lingbo Chenga and Mahdi Tavakolia. "Neural-Network-Based Heart Motion Prediction for Ultrasound-Guided Beating-Heart Surgery". In: *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*. IEEE. 2019, pp. 437–442.

[162] Hui Lin et al. "Towards real-time respiratory motion prediction based on long short-term memory neural networks". In: *Physics in Medicine & Biology* 64.8 (2019), p. 085010.

[163] Troy P Teo et al. "Feasibility of predicting tumor motion using online data acquired during treatment and a generalized neural network optimized with offline patient tumor trajectories". In: *Medical physics* 45.2 (2018), pp. 830–845.

[164] Qiaoying Huang et al. "Dynamic MRI Reconstruction with Motion-Guided Network". In: *MIDL*. 2019.

[165] Qiaoying Huang et al. "Dynamic MRI reconstruction with end-to-end motion-guided network". In: *Medical Image Analysis* 68 (2021), p. 101901.

[166] Majid Mafi and Saeed Montazeri Moghadam. "Real-time prediction of tumor motion using a dynamic neural network". In: *Medical & Biological Engineering & Computing* (2020), pp. 1–11.

[167] Dongyeon Lee et al. "Four-Dimensional CBCT Reconstruction Based on a Residual Convolutional Neural Network for Improving Image Quality". In: *Journal of the Korean Physical Society* 75.1 (2019), pp. 73–79.

[168] Hui Lin et al. “A Super-Learner Model for tumor Motion prediction and Management in Radiation therapy: Development and feasibility evaluation”. In: *Scientific reports* 9.1 (2019), pp. 1–11.

[169] You Zhang, Xiaokun Huang, and Jing Wang. “Advanced 4-dimensional cone-beam computed tomography reconstruction by combining motion estimation, motion-compensated reconstruction, biomechanical modeling and deep learning”. In: *Visual Computing for Industry, Biomedicine, and Art* 2.1 (2019), pp. 1–15.

[170] Zhehao Zhang et al. “Deep learning-based motion compensation for four-dimensional cone-beam computed tomography (4D-CBCT) reconstruction”. In: *Medical physics* 50.2 (2023), pp. 808–820.

[171] Andrea Mendizabal, Pablo Márquez-Neila, and Stéphane Cotin. “Simulation of hyperelastic materials in real-time using deep learning”. In: *Medical image analysis* 59 (2020), p. 101569.

[172] Lijuan Shi et al. “Respiratory Prediction Based on Multi-Scale Temporal Convolutional Network for Tracking Thoracic Tumor Movement”. In: *Frontiers in Oncology* 12 (2022). ISSN: 2234-943X. DOI: 10.3389/fonc.2022.884523. URL: https://www.frontiersin.org/articles/10.3389/fonc.2022.884523.

[173] Guangjun Li et al. “Machine learning for predicting accuracy of lung and liver tumor motion tracking using radiomic features”. In: *Quantitative Imaging in Medicine and Surgery* 12 (2023). DOI: 10.21037/qims-22-621.

[174] Jun Lv et al. “Respiratory motion correction for free-breathing 3D abdominal MRI using CNN-based image registration: a feasibility study”. In: *The British journal of radiology* 91.xxxx (2018), p. 20170788.

[175] Hessam Sokooti et al. “Nonrigid image registration using multi-scale 3D convolutional neural networks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 232–239.

[176] Koen AJ Eppenhof and Josien PW Pluim. “Pulmonary CT registration through supervised learning with convolutional neural networks”. In: *IEEE transactions on medical imaging* 38.5 (2018), pp. 1097–1105.

[177] Hristina Uzunova et al. “Training CNNs for image registration from few samples with model-based data augmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 223–231.

[178] Alina Giger et al. “Respiratory motion modelling using cGANs”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*. Springer. 2018, pp. 81–88.

[179] Yabo Fu et al. “LungRegNet: an unsupervised deformable image registration method for 4D-CT lung”. In: *Medical Physics* 47.4 (2020), pp. 1763–1774.

[180] Richard Castillo et al. “A reference dataset for deformable image registration spatial accuracy evaluation using the COPDgene study archive”. In: *Physics in Medicine & Biology* 58.9 (2013), p. 2861.

[181] Hessam Sokooti et al. “3D Convolutional Neural Networks Image Registration Based on Efficient Supervised Learning from Artificial Deformations”. In: *arXiv preprint arXiv:1908.10235* (2019).

[182] Thilo Sentker, Frederic Madesta, and René Werner. “GDL-FIRE 4D: Deep Learning-Based Fast 4D CT Image Registration”. In: *International Conference*

on Medical Image Computing and Computer-Assisted Intervention. Springer. 2018, pp. 765–773.

[183] Chen Qin et al. "Biomechanics-Informed Neural Networks for Myocardial Motion Tracking in MRI". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by Anne L. Martel et al. Cham: Springer International Publishing, 2020, pp. 296–306. ISBN: 978-3-030-59716-0.

[184] You Zhang. "An unsupervised 2D–3D deformable registration network (2D3D-RegNet) for cone-beam CT estimation". In: *Physics in Medicine & Biology* 66.7 (2021), p. 074001.

[185] Filip Jacobs et al. "A fast algorithm to calculate the exact radiological path through a pixel or voxel space". In: *Journal of computing and information technology* 6.1 (1998), pp. 89–94.

[186] Robert L Siddon. "Fast calculation of the exact radiological path for a three-dimensional CT array". In: *Medical physics* 12.2 (1985), pp. 252–255.

[187] Hua-Chieh Shao et al. "Real-time MRI motion estimation through an unsupervised k-space-driven deformable registration network (KS-RegNet)". In: *Physics in Medicine & Biology* 67.13 (2022), p. 135012.

[188] Brian C. Lee et al. "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)*. 2022, pp. 1–5. DOI: 10.1109/ISBI52829.2022.9761705.

[189] François Lecomte, Jean-Louis Dillenseger, and Stéphane Cotin. "CNN-based real-time 2D-3D deformable registration from a single X-ray projection". In: *arXiv preprint arXiv:2212.07692* (2022).

[190] Huiqiao Xie et al. “Inter-fraction deformable image registration using unsupervised deep learning for CBCT-guided abdominal radiotherapy”. In: *Physics in Medicine and Biology* (2023).

[191] Guoya Dong et al. “2D/3D Non-Rigid Image Registration via Two Orthogonal X-ray Projection Images for Lung Tumor Tracking”. In: *Bioengineering* 10.2 (2023), p. 144.

[192] Jingjing Dai et al. “Volumetric tumor tracking from a single cone-beam X-ray projection image enabled by deep learning”. In: *Medical Image Analysis* (2023), p. 102998. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2023.102998. URL: https://www.sciencedirect.com/science/article/pii/S136184152300258X.

[193] Taesung Park et al. “Contrastive learning for unpaired image-to-image translation”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer. 2020, pp. 319–345.

[194] Salim Balik et al. “Evaluation of 4-dimensional computed tomography to 4-dimensional cone-beam computed tomography deformable image registration for lung cancer adaptive radiation therapy”. In: *International Journal of Radiation Oncology* Biology* Physics* 86.2 (2013), pp. 372–379.

[195] Kenneth Clark et al. “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository”. In: *Journal of digital imaging* 26 (2013), pp. 1045–1057.

[196] Geoffrey D Hugo et al. “Data from 4D lung imaging of NSCLC patients”. In: *The Cancer Imaging Archive* 10 (2016), K9.

[197] Geoffrey D Hugo et al. “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 (2017), pp. 762–771.

[198] Jamie R McClelland et al. “Respiratory motion models: a review”. In: *Medical image analysis* 17.1 (2013), pp. 19–42.

[199] Jamie R McClelland et al. “A generalized framework unifying image registration and respiratory motion models and incorporating image reconstruction, for partial image data or full images”. In: *Physics in Medicine & Biology* 62.11 (2017), p. 4273.

[200] Minghao Guo et al. “Reconstruction of a high-quality volumetric image and a respiratory motion model from patient CBCT projections”. In: *Medical physics* 46.8 (2019), pp. 3627–3639.

[201] Wendy Harris et al. “A technique for generating volumetric cine-magnetic resonance imaging”. In: *International Journal of Radiation Oncology* Biology* Physics* 95.2 (2016), pp. 844–853.

[202] Elena H Tran et al. “Evaluation of MRI-derived surrogate signals to model respiratory motion”. In: *Biomedical physics & engineering express* 6.4 (2020), p. 045015.

[203] Liset Vázquez Romaguera et al. “Probabilistic 4D predictive model from in-room surrogates using conditional generative networks for image-guided radiotherapy”. In: *Medical image analysis* 74 (2021), p. 102250.

[204] Tal Mezheritsky et al. “Population-based 3D respiratory motion modelling from convolutional autoencoders for 2D ultrasound-guided radiotherapy”. In: *Medical Image Analysis* 75 (2022), p. 102260.

[205] Cong Liu et al. “NuTracker: a coordinate-based neural network representation of lung motion for intrafraction tumor tracking with various surrogates in radiotherapy”. In: *Physics in Medicine & Biology* 68.1 (Dec. 2022), p. 015006. DOI: 10.1088/1361-6560/aca873. URL: https://dx.doi.org/10.1088/1361-6560/aca873.

[206] Nanyang Wang et al. “Pixel2mesh: Generating 3d mesh models from single rgb images”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 52–67.

[207] Edward J Smith et al. “GEOMetrics: Exploiting geometric structure for graph-encoded objects”. In: *arXiv preprint arXiv:1901.11461* (2019).

[208] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. “Convolutional mesh regression for single-image human shape reconstruction”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4501–4510.

[209] Sizhuo Zhang and Nanfeng Xiao. “Detailed 3D Human Body Reconstruction From a Single Image Based on Mesh Deformation”. In: *IEEE Access* 9 (2021), pp. 8595–8603. DOI: 10.1109/access.2021.3049548. URL: https://doi.org/10.1109%5C%2Faccess.2021.3049548.

[210] Shuqiong Wu et al. “Reconstructing 3D Lung Shape from a Single 2D Image during the Deaeration Deformation Process using Model-based Data Augmentation”. In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, May 2019. DOI: 10.1109/bhi.2019.8834454. URL: https://doi.org/10.1109%5C%2Fbhi.2019.8834454.

[211] Xingde Ying et al. “X2CT-GAN: reconstructing CT from biplanar X-rays with generative adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019, pp. 10619–10628.

[212] Ran Wei et al. “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 (2020), p. 065012.

[213] Yifan Wang, Zichun Zhong, and Jing Hua. “DeepOrganNet: On-the-Fly Reconstruction and Visualization of 3D / 4D Lung Models from Single-View Projections by Deep Deformation Network”. In: *IEEE Transactions on Visualization and Computer Graphics* (2019), pp. 1–1. DOI: 10.1109/tvcg.2019.2934369. URL: https://doi.org/10.1109%5C%2Ftvcg.2019.2934369.

[214] Andrew G Howard et al. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. In: *arXiv preprint arXiv:1704.04861* (2017).

[215] Fei Tong et al. “X-ray2Shape: Reconstruction of 3D Liver Shape from a Single 2D Projection Image”. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. 2020, pp. 1608–1611. DOI: 10.1109/EMBC44109.2020.9176655.

[216] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[217] Megumi Nakao et al. “Image-to-Graph Convolutional Network for Deformable Shape Reconstruction from a Single Projection Image”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 259–268.

[218] Megumi Nakao, Mitsuhiro Nakamura, and Tetsuya Matsuda. “Image-to-Graph Convolutional Network for 2D/3D Deformable Model Registration of Low-

Contrast Organs". In: *IEEE Transactions on Medical Imaging* 41.12 (2022), pp. 3747–3761. DOI: 10.1109/TMI.2022.3194517.

[219] Shaolin Lu et al. "Prior information-based high-resolution tomography image reconstruction from a single digitally reconstructed radiograph". In: *Physics in Medicine & Biology* 67.8 (Apr. 2022), p. 085004. DOI: 10.1088/1361-6560/ac508d. URL: https://doi.org/10.1088/1361-6560/ac508d.

[220] Hua-Chieh Shao et al. "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 (2022), p. 115009.

[221] Hua-Chieh Shao et al. "Real-time liver tumor localization via combined surface imaging and a single x-ray projection". In: *Physics in Medicine & Biology* 68.6 (2023), p. 065002.

[222] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016).

[223] Petar Veličković et al. "Graph attention networks". In: *arXiv preprint arXiv:1710.10903* (2017).

[224] Petar Veličković et al. "Graph Attention Networks". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=rJXMpikCZ.

[225] Yuxin Wu and Kaiming He. "Group normalization". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[226] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

[227] Takeru Miyato et al. “Spectral Normalization for Generative Adversarial Networks”. In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=B1QRgziT-.

[228] Jamie R McClelland et al. *SuPReMo*. 2017. URL: https://github.com/UCL/SuPReMo.

[229] Simon Rit et al. “The Reconstruction Toolkit (RTK), an open-source cone-beam CT reconstruction toolkit based on the Insight Toolkit (ITK)”. In: *Journal of Physics: Conference Series*. Vol. 489. 1. IOP Publishing. 2014, p. 012079.

[230] S Veloza, HU Kauczor, and W Stiller. “Performance of Attenuation-based Dynamic CT Beam-shaping Filtration for Elliptical Subject Geometries in Dependence of Fan-and Projection-angle”. In: *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada*. Springer. 2015, pp. 95–98.

[231] Jun-Yan Zhu et al. “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2242–2251. DOI: 10.1109/ICCV.2017.244.

[232] Karel Zuiderveld. “Contrast limited adaptive histogram equalization”. In: *Graphics gems IV*. 1994, pp. 474–485.

[233] Hao-Chun Lu, El-Wui Loh, and Shih-Chen Huang. “The classification of mammogram using convolutional neural network with specific image preprocessing for breast cancer detection”. In: *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE. 2019, pp. 9–12.

[234] Andrew Aitken et al. “Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize”. In: *arXiv preprint arXiv:1707.02937* (2017).

[235] Qianqian Fang and David A Boas. “Tetrahedral mesh generation from volumetric binary and grayscale images”. In: *2009 IEEE international symposium on biomedical imaging: from nano to macro*. Ieee. 2009, pp. 1142–1145.

[236] David Ahmedt-Aristizabal et al. “Graph-based deep learning for medical diagnosis and analysis: past, present and future”. In: *Sensors* 21.14 (2021), p. 4758.

[237] Nanyang Wang et al. “Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images”. In: *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 55–71. DOI: 10.1007/978-3-030-01252-6_4. URL: https://doi.org/10.1007%5C%2F978-3-030-01252-6_4.

[238] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[239] Matthias Fey and Jan E. Lenssen. “Fast Graph Representation Learning with PyTorch Geometric”. In: *ICLR Workshop on Representation Learning on Graphs and Manifolds*. 2019.

[240] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[241] Paul Viola and William M Wells III. “Alignment by maximization of mutual information”. In: *International journal of computer vision* 24.2 (1997), pp. 137–154.

[242] Jonghye Woo, Maureen Stone, and Jerry L Prince. “Multimodal registration via mutual information incorporating geometric and spatial context”. In: *IEEE Transactions on Image Processing* 24.2 (2014), pp. 757–769.

[243] Debapriya Sengupta, Phalguni Gupta, and Arindam Biswas. “A survey on mutual information based medical image registration algorithms”. In: *Neurocomputing* 486 (2022), pp. 174–188.

[244] Han Zhang et al. “Self-Attention Generative Adversarial Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 7354–7363. URL: https://proceedings.mlr.press/v97/zhang19d.html.

[245] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[246] Arezoo Zakeri et al. “4D-Precise: Learning-Based 3D Motion Estimation and High Temporal Resolution 4DCT Reconstruction from In-Treatment 2D t X-Ray Projections”. In: *Available at SSRN 4661696* ().

[247] Arezoo Zakeri et al. “DragNet: Learning-based deformable registration for realistic cardiac MR sequence generation from a single frame”. In: *Medical Image Analysis* 83 (2023), p. 102678. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2022.102678. URL: https://www.sciencedirect.com/science/article/pii/S1361841522003061.