Imaging Error Reduction for MR Guided Radiotherapy with Deep Learning-Based Intra-Frame Motion Compensation

Zhuojie Sui

München 2025

Imaging Error Reduction for MR Guided Radiotherapy with Deep Learning-Based Intra-Frame Motion Compensation

Zhuojie Sui

DISSERTATION
an der Fakultät für Physik
der Ludwig-Maximilians-Universität
München

vorgelegt von
Zhuojie Sui
aus Yantai, China

München, den 08.05.2025

Erstgutachter: Prof. Dr. Marco Riboldi
Zweitgutachter: Prof. Dr. Chiara Paganelli
Tag der mündlichen Prüfung: 10.07.2025

Zusammenfassung

Die Strahlentherapie in Anwesenheit intra-fraktioneller Bewegung kann erheblich von der Echtzeit-Bildgebung mit Magnetresonanztomographie (MRT) profitieren, da diese eine überlegene Weichteilkontrastierung bietet und keine ionisierende Strahlung verwendet. Bewegungsbedingte Bildgebungsfehler wurden jedoch als Hauptursache für die gesamte Schleifenlatenz in der MRT-geführten Strahlentherapie (MRgRT) identifiziert. Diese Fehler führen zu verbleibenden geometrischen Verfolgungsfehlern und beeinträchtigen somit die Wirksamkeit des aktiven Bewegungsmanagements. In dieser Dissertation wird die Möglichkeit untersucht, diese Fehler in der MRgRT durch Deep-Learning-basierte intra-frame Bewegungskompensation zu reduzieren.

Zunächst wurde ein bewegungsabhängiges k-Raum-Simulationsverfahren entwickelt, um das Verhalten der dynamischen MRT-Bildgebung sowie bewegungsbedingte Bildfehler zu untersuchen. Darauf aufbauend wurde eine Methodik zur Erstellung und Erweiterung von intra-frame Bewegungsdatensätzen vorgeschlagen, bei der bewegungsverfälschte Daten mit ihren Echtzeit-Ground-Truth-Pendants kombiniert wurden, wobei der Schwerpunkt auf schnellen anatomischen Veränderungen lag. Konkret wurden auf der Grundlage einer groß-zu-feinen gitterbasierten Repräsentation patientenspezifischer Bewegungsdaten digitale 4D-MRT-Phantome zur Modellierung von Lungenkrebspatienten erzeugt, und ein spezielles intra-frame Bewegungsmodell wurde mittels stückweiser linearer Approximation zwischen aufeinanderfolgenden Kontrollpunkten aufgebaut. Zusätzlich wurde ein Verfahren zur Erzeugung von Abweichungen von Bewegungsmustern eingeführt, um potenzielle Positionen anatomischer Strukturen umfassend zu erforschen und die Vielfalt intra-frame Bewegungsverläufen zu erhöhen.

Zweitens wurde eine Machbarkeitsstudie mit kartesischer Cine-MRT durchgeführt, die zeigte, dass UNet-Modelle intra-frame Bewegungen wirksam kompensieren können, indem sie das Bild an der Endposition der Aufnahme aus bewegungsverfälschten Eingangsdaten schätzen. Quantitativ stieg im Testdatensatz für die Konturierung des makroskopischen Tumorvolumens (GTV) der mediane Dice Similarity Coefficient (DSC) von 89% auf 97%, während der 95. Perzentilwert der Hausdorff-Distanz (HD_{95}) von 4,1 mm auf 1,4 mm sank. Geometrische Fehler in Zielstrukturen mit ausgeprägten intra-frame Deformationen konnten erfolgreich korrigiert werden und zeigten eine enge Übereinstimmung mit dem Ground Truth hinsichtlich Form und Position des Zielvolumens. Die Saliency Maps wiesen darauf hin, dass sich das Modell bei der Inferenz hauptsächlich auf die später erfassten k-Raum-Komponenten konzentrierte

und entsprechend im Ortsraum auf die Ränder der sich bewegenden Strukturen an deren Echtzeit-Endposition.

Drittens wurde eine Machbarkeitsstudie mit radialer Cine-MRT durchgeführt, in der "TransSin-UNet" vorgestellt wurde – ein neuartiges Deep-Learning-Framework im Dual-Domain-Ansatz. Innerhalb des radialen k-Raum-Rekonstruktionsfensters wurden die weit reichenden räumlich-zeitlichen Abhängigkeiten in der Sinogramm-Darstellung der Speichen durch ein Transformer-Encoder-Subnetzwerk modelliert, gefolgt von einem UNet-Subnetzwerk im Ortsraum zur Verfeinerung auf Pixelebene. Das Netzwerk wurde auf Datensätzen mit unterschiedlichen azimutalen Inkrementen der radialen Profile trainiert und umfassend evaluiert. Im Vergleich zur konventionellen direkten Bildrekonstruktion erforderte TransSin-UNet nur zusätzliche 4,8 ms pro Bild zur Kompensation von bewegungsverfälschten Speichen. Es übertraf konsistent Architekturen, die ausschließlich auf Transformer-Encodern oder UNets basierten, in sämtlichen Vergleichsstudien und führte zu einer deutlichen Verbesserung der Bildqualität und Zielpositionierungsgenauigkeit. Der normalisierte Root Mean Squared Error (NRMSE) sank um 50 % vom ursprünglichen Mittelwert von 0,188, während der mittlere DSC des GTV in den untersuchten Testfällen von 85,1 % auf 96,2 % anstieg. Darüber hinaus konnten die Ground-Truth-Positionen anatomischer Strukturen mit ausgeprägten Deformationen präzise bestimmt werden.

Diese Arbeit stellt einen bedeutenden Fortschritt auf dem Weg zur klinischen Umsetzung von Strategien zur Reduktion von Tracking-Fehlern in Cine-MRT dar und unterstützt ein verbessertes Echtzeit-Bewegungsmanagement in der MRgRT.

Abstract

Radiotherapy in the presence of intra-fractional motion can significantly benefit from real-time magnetic resonance imaging (MRI) guidance, owing to its superior soft tissue contrast and the absence of ionizing radiation. However, motion-related imaging errors have been identified as the primary contributor to overall loop latency in MR-guided radiotherapy (MRgRT), leading to residual geometric tracking errors and subsequently affecting the effectiveness of active motion management. This thesis explores the feasibility of reducing these errors in MRgRT through deep learning-based intra-frame motion compensation techniques.

Firstly, a motion-dependent k-space sampling simulation procedure was developed to investigate dynamic MR imaging behavior and motion-related imaging errors. Building upon this, a methodology for intra-frame motion dataset creation and augmentation was proposed, pairing the motion-corrupted data with its real-time ground-truth counterpart, with a primary focus on rapid anatomical changes. Specifically, based on a coarse-to-fine grid-scale representation of patient-specific motion data, 4D MRI digital anthropomorphic phantoms were generated to model lung cancer patients, and a dedicated intra-frame motion model was constructed using a piecewise linear approximation between consecutive control points. Additionally, a motion pattern perturbation scheme was introduced to comprehensively explore potential anatomical structure positions and enhance the diversity of intra-frame motion trajectories.

Secondly, a proof-of-concept study in Cartesian cine-MRI was conducted, demonstrating that UNet models can effectively compensate for intra-frame motion by estimating the final-position image at the end of frame acquisition from motion-corrupted input. Quantitatively, in the testing dataset for gross tumor volume (GTV) contouring, the median Dice similarity coefficient (DSC) increased from 89% to 97%, while the 95th percentile Hausdorff distance (HD_{95}) decreased from 4.1 mm to 1.4 mm. Geometric errors in targets undergoing considerable intra-frame deformations were successfully corrected, exhibiting close agreement with the ground truth in terms of both target shape and position. The saliency maps indicated that the model predominantly focused on the later-acquired k-space components for inference and, correspondingly in the spatial domain, the edges of the moving structures at their real-time final positions.

Thirdly, a proof-of-concept study in radial cine-MRI was conducted, proposing "TransSin-UNet", a novel dual-domain deep learning framework. Within the radial k-space reconstruction window, the long-distance spatial-temporal dependencies among the sinogram representation of the spokes were modeled by a transformer encoder sub-

network, followed by a UNet subnetwork operating in the spatial domain for pixel-level refinement. The network was trained and extensively evaluated across datasets with varying azimuthal radial profile increments. TransSin-UNet required only an additional 4.8 ms per frame for compensation compared to conventional direct image reconstruction using motion-corrupted spokes. It consistently outperformed architectures relying solely on transformer encoders or UNets across all comparative evaluations, leading to a noticeable enhancement in image quality and target positioning accuracy. The normalized root mean squared error (NRMSE) decreased by 50% from the initial average of 0.188, whereas the mean DSC of GTV increased from 85.1% to 96.2% in the investigated testing cases. Furthermore, the ground-truth positions of anatomical structures experiencing substantial deformations were precisely derived.

This work constitutes a substantial advancement toward the clinical implementation of cine-MR tracking error reduction strategies to support enhanced real-time motion management in MRgRT.

Contents

List of Figures

List of Tables

List of Abbreviations

AAC	Amplitude amplification coefficient
AAPM	American Association of Physicists in Medicine
ACS	Auto-calibration signal
AI	Artificial intelligence
AP	Anterior-posterior
CBCT	Cone beam computed tomography
CNN	Convolutional neural network
COM	Center of mass
CS	Compressed sensing
DC	Direct current
DCE	Dynamic contrast-enhanced
DCF	Density correction factor
DFT	Discrete Fourier transform
DIR	Deformable image registration
DNA	Deoxyribonucleic acid
DSC	Dice similarity coefficient
DVF	Displacement vector field
DW	Diffusion-weighted
EBRT	External beam radiation therapy
EFE	Electron focusing effect
ERE	Electron return effect
FFT	Fast Fourier transform
FID	Free induction decay
FPS	Frames per second
GE	Gradient echo
GRAPPA	Generalized autocalibrating partially parallel acquisition
GTV	Gross tumor volume
HD_{95}	95th percentile Hausdorff distance
HFC	Higher frequency component
ICRU	Radiation units and measurements
IEC	International Electrotechnical Commission
IFT	Inverse Fourier transform
IGRT	Image-guided radiation therapy
IFFT	Inverse fast Fourier transform

IMRT	Intensity-modulated radiation therapy
IQR	Interquartile range
ITV	Internal target volume
kV	Kilovoltage
LFC	Lower-frequency component
LR	Left-right
MAE	Mean absolute error
MLC	Multi-leaf collimator
MRI	Magnetic resonance imaging
MRgRT	MR-guided radiotherapy
MRiPT	MR-integrated proton therapy
MSE	Mean squared error
MTF	Modulation transfer function
MV	Megavolts
NMR	Nuclear magnetic resonance
NRMSE	Normalized root mean squared error
NTCP	Normal tissue complication probability
NUFFT	Non-uniform fast Fourier transform
OAR	Organs at risk
OOD	Out-of-distribution
PTV	Planning target volume
RC-4D-MRI	Respiratory-correlated 4D-MRI
RF	Radio frequency
RT	Radiation therapy
RT-4D-MRI	Real-time 4D-MRI
SBRT	Stereotactic body radiotherapy
SE	Spin echo
SENSE	Sensitivity encoding
SI	Superior-inferior
SinTE	Sinogram transformer encoder
SSIM	Structural similarity
TCP	Tumor control probability
TG	Task group
TrueFISP	True fast imaging with steady-state precession
VMAT	Volumetric-modulated arc therapy

Chapter 1

INTRODUCTION AND MOTIVATION

1.1 Radiotherapy

Cancer ranks as the second leading cause of death globally, following cardiovascular diseases, responsible for approximately 9.6 million deaths, or one in six deaths, in 2018 [1]. It is characterized by the uncontrolled proliferation of tumor cells, caused by defects in the cellular reproduction cycle, which invades nearby tissues and potentially spreads to distant places in the body, a process known as metastasis [2].

The current treatment of cancer encompasses both isolated or combined modalities, with the three main ones being surgery, systemic therapy (such as chemotherapy or hormonal therapy), and radiation therapy (RT). It is widely reported that around a quarter of all cancer patients ultimately receive RT, while recommendations aimed at enhancing overall survival suggest increasing this proportion to fifty percent, with external beam radiation therapy (EBRT) recognized as the best practice care in about half of all cancer cases. [3–5]. This thesis will focus exclusively on radiotherapy, specifically EBRT, as a non-invasive and non-pharmacological method to target and eradicate tumor cells.

The primary mechanism through which radiation kills cells is by causing double-strand breaks in the deoxyribonucleic acid (DNA), which are challenging for the cell to repair and may result in its inability to replicate. The effectiveness of radiation therapy relies on the different response of malignant and normal tissues to ionizing radiation exposure, characterized by the tumor control probability (TCP) and normal tissue complication probability (NTCP), respectively [6]. The difference between TCP and NTCP as a function of the delivered dose defines a therapeutic window [7], as illustrated in Fig. 1.1, facilitating the prescription of an optimal radiation dose that maximizes the likelihood of tumor control while minimizing toxicity to surrounding healthy tissues to acceptable levels [8].

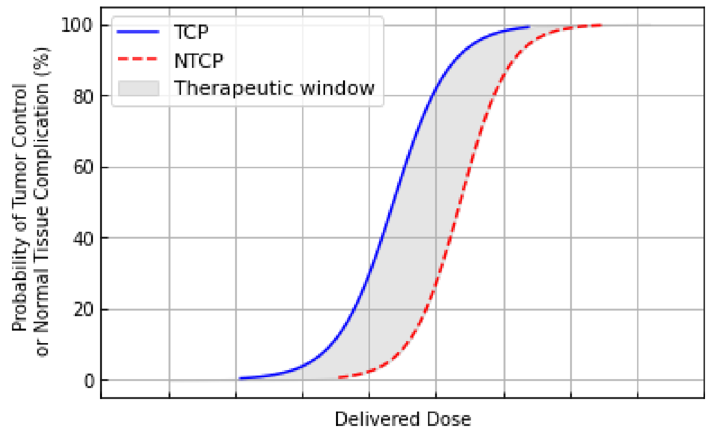

Figure 1.1: Schematic representation of the therapeutic window. The dose response curves of TCP and NTCP are modeled using sigmoid functions.

The narrow therapeutic window places stricter demands on the delivery accuracy of a prescribed dose, as dosimetric uncertainties lead to either reductions in TCP or increases in NTCP relative to the optimized expected value, both of which worsen the clinical outcome. Notably, at the steepest portions of the the most critical dose-response curves, a 5% variation in dose can produce 10-20% changes in TCP and 20-30% changes in NTCP [9]. For external photon beam radiotherapy, the International Commission on Radiation Units and Measurements (ICRU) Report 50 recommends a target dose uniformity within +7% and -5% of the dose delivered to a well-defined prescription point within the target [10, 11].

1.2 Motion management and real-time motion monitoring

Geometric uncertainties translate into dosimetric uncertainties, resulting in potential underdosage of the target region and/or overdosage in nearby organs at risk (OAR), making them critical considerations in RT. These uncertainties can stem from various factors, including treatment machine specifications and tolerances, simulation and treatment setup, anatomical alterations between fractions (inter-fractional variations) [12] and shorter-term patient or organ motion during a treatment session (intra-fractional motion) [9]. With the advancement of delivery techniques in EBRT, such as intensity-modulated radiation therapy (IMRT) [13] and volumetric-modulated arc therapy (VMAT) [14], which are designed to achieve highly conformal dose distributions shaped around the planning target volume (PTV), accurate target and OAR localization becomes even more relevant [9].

Image-guided radiation therapy (IGRT) [15] has become a cornerstone of modern precision radiation oncology, playing a crucial role in reducing geometric uncertainties, particularly those introduced by patient positioning and inter-/intra-fraction motion. With the application of advanced in-room imaging, patient setup can be verified and adjusted prior to each fraction to ensure that the target volume aligns correctly with the treatment-planning position. Moreover, the baseline treatment plan can be re-optimized to adapt to the daily anatomical-pathological situation in the treatment position, effectively accounting for the inter-fractional changes [16]. Furthermore, advancements in real-time imaging of moving targets provide a foundation for developing strategies to manage intra-fractional motion during irradiation [17], such as breathing and heartbeat, spontaneous motion [18], and baseline drifts [19].

As highlighted in the report of the American Association of Physicists in Medicine (AAPM) Task Group (TG) 76 [20], intra-fractional motion is an issue of growing significance in the era of IGRT. Certain types of motion, particularly respiratory motion, can be patient-specific, difficult to predict, irregular, and vary over time. Additionally, the motion variations associated with tumor location and pathology result in distinct individual patterns in displacement, direction, and motion phase. Their assessment and accommodation are therefore of critical importance.

In clinical practice, techniques for intra-fractional motion management are generally classified into passive and active approaches [21, 22]. Margins are a widely employed passive approach aiming at ensuring target coverage in the presence of intra-fractional motion. This can involve defining an internal target volume (ITV) that encompasses the full extent of tumor motion as observed in the treatment-planning stage, or applying a statistical margin recipe, such as the mid-ventilation approach [23–25]. Nevertheless, these approaches often result in larger irradiated volumes, subjecting close-by OARs to higher doses [17, 26], and may still fail to provide adequate target coverage, particularly when tumor drift occurs. By contrast, active real-time motion management approaches, including gating and tracking, offer enhanced targeting accuracy and facilitate a safe margin reduction [27–29]. Gating activates the beam only when the target moves inside a predefined boundary [30], while tracking ensures continuous synchronization between the beam and the moving target [31–33]. When comparing these active approaches in the context of respiratory motion management, gating during free-breathing decreases the duty cycle while gating in breath-hold requires patient compliance. Tracking, on the other hand, is more efficient but involves greater technical complexity and is currently limited to specialized commercial platforms [34].

Extensive studies have reported convincing evidence in favor of active motion management from the perspectives of geometric accuracy, dosimetric precision, and clinical outcomes [21, 27, 35–37]. The AAPM TG76 report recommends implementing active motion management when respiratory motion exceeds an amplitude of 5 mm, if it can significantly enhance OAR sparing, or when necessary to meet clinical objectives [20].

Real-time motion monitoring is essential for active motion management to trigger the beam on/off signal during gating or maintain continuous beam-target realignment in the tracking feedback chain. Additionally, real-time motion monitoring is particularly crucial for stereotactic body radiotherapy (SBRT) [38], which delivers highly collimated beams to the lesion at significantly higher doses and with much greater precision than traditional EBRT, necessitating tight margins for OAR sparing [21]. Furthermore, time-resolved motion monitoring data can be utilized to estimate accurate dose accumulation for each fraction [39, 40], thereby facilitating treatment adaptation if the tumor coverage is inadequate or OAR constraints are violated [41, 42].

The long-term goal of the RT community to 'see what we treat, as we treat' and adapt treatment in real-time has driven the advancement and widespread implementation of numerous online motion monitoring and mitigation techniques [17]. Infrared-based or optical surface monitoring [43, 44], kilovoltage (kV) or megavoltage (MV) X-ray imaging [45, 46], magnetic resonance imaging (MRI) [32, 47], etc., have been extensively integrated into treatment delivery devices, such as linear accelerators (linac), and are now routinely utilized in clinical practice. For a more comprehensive review and comparison of the current real-time intra-fractional motion monitoring techniques in EBRT, please refer to [17].

1.3 MRgRT

Cone beam computed tomography (CBCT) is a widely used imaging modality, which has increasingly become the standard method for IGRT in recent years [48,49]. Nonetheless, CBCT presents several inherent shortcomings. First, the poor soft tissue contrast [50] makes it challenging to distinguish tumors from surrounding tissues. Second, CBCT produces suboptimal image quality. The area detector captures scattered radiation from all directions, with nonlinear attenuation further contributing to image degradation and increased noise [51]. Third, while CBCT generally delivers lower radiation doses than conventional CT, the additional imaging dose [52] to radiosensitive organs remains a consideration, potentially leading to side effects. This is particularly concerning for real-time intra-fractional motion monitoring. Fourth, CBCT can significantly underestimate target motion ranges, raising concerns about its suitability for motion management [16, 53].

In light of these limitations, MRI, with its high soft tissue contrast, absence of ionizing radiation, functional imaging capabilities, and versatile modalities, emerges as an ideal alternative for implementing IGRT. MR-guided radiotherapy (MRgRT) is widely regarded as a game changer for numerous tumor sites [54], marking a new era of precision treatment. The superior soft tissue contrast of MRI significantly enhances delineation precision during treatment planning and holds great potential for improving localization accuracy of moving targets during beam delivery. The dose-free nature of MRI allows for frequent verification of treatment adaptation strategies and continuous, long-term monitoring of intra-fractional anatomical variations. Additionally, functional quantitative MRI techniques [55], such as dynamic contrast-enhanced (DCE) and diffusion-weighted (DW) MRI [56], integrated into multi-parametric analyses, have the potential to enhance the entire RT workflow [57]. These contributions span diagnosis [58, 59], contouring [60], dose optimization [61], treatment monitoring [62], and response assessment [63], thereby advancing treatment personalization [64].

Over the past decade, substantial research and commercial efforts have been dedicated to integrating onboard MR scanners with treatment units [30, 65–69]. Table 1.1 summarizes the existing MRgRT approaches employing linear accelerators (MR-Linac systems), which feature varying configurations regarding magnetic field strength, radiation source and energy, as well as the orientation of the static magnetic field relative to the radiation beam [70]. Among these, the ViewRay MRIdian [71], Elekta Unity [72], and Aurora-RT [73] are currently available for commercial use. The world's inaugural MRgRT treatment was carried out with the Cobalt-60-based MRIdian system in 2014 [74], followed by the first MRI-Linac patient treatment utilizing Unity in 2017 [75].

Table 1.1: Configurations of representative MR-Linac systems. The data were compiled from seminal publications in the field.

System	Company/Institute*	Radiation source	Field orientation	Field strength
MRIdian	ViewRay	^{60}Co / 6 MV	perpendicular	0.35 T
Unity	Elekta	7 MV	perpendicular	1.5 T
Aurora-RT	MagnetTx	4/6 MV	parallel	0.56 T
Australia	Ingham*	6 MV	parallel/perpendicular	1.0 T

Real-time monitoring of tumor and OAR motion in today's clinical MR-Linac systems is achieved through online 2D+t cine-MR imaging. During irradiation, cine-MR frames are continuously acquired with rapid imaging sequences, such as balanced Steady State Free Precession (bSSFP) [76] and spoiled gradient echo [77], at frame rates of a few Hz. In the clinical setup of the ViewRay MRIdian MR-Linac system, the bSSFP sequence employing a Cartesian k-space readout trajectory achieves a temporal resolution of 4 Hz, while a radial k-space readout variant provides an enhanced temporal resolution of 8 frames per second (FPS).

A 4D-CT scan of the moving anatomy is typically acquired to evaluate the extent of respiratory-induced motion, forming a key component of conventional radiotherapy treatment planning for the thoracic and abdominal regions, such as lung tumors. However, due to challenges like the inherent trade-off between spatial and temporal resolution, 4D-MRI—including the respiratory-correlated 4D-MRI (rc-4D-MRI) and real-time 4D-MRI (rt-4D-MRI)—is not yet offered by MR-Linac vendors [54]. Despite this, interest in both approaches has grown steadily in recent years due to their potential applications in MRgRT [78, 79]. Specifically, rc-4D-MRI holds promise for improving treatment planning, whereas rt-4D-MRI could enhance real-time target and OAR localization during beam delivery, particularly when significant out-of-plane motion is present in 2D+t cine-MRI [16].

The vendor's cine-MRI data are highly effective in facilitating active intra-fractional motion management, with clinical studies already published [32, 80–84]. Gating treatment for mobile targets has become a routine practice in clinical applications on the ViewRay MRIdian MR-Linac. A gating boundary is defined prior to treatment as an expansion of the target contour. During irradiation, the system employs an optical flow deformable image registration (DIR) algorithm [85] to deform the target contour from the reference image to each cine-MR frame, enabling real-time localization of the target position. The relative overlap between the real-time target contour and the gating boundary is then evaluated, referred to as the target out percentage, and compared to a predefined threshold, typically set between 5% and 10%. Beam delivery occurs only when the target out percentage remains below the threshold (classified as target in); otherwise, the beam is automatically paused. The Elekta Unity MR-Linac facilitates multi-leaf collimator (MLC)-tracking for moving targets. The system reshapes and repositions the radiation beam using a 160-leaf MLC with optically encoded leaf positions. While the treatment head remains stationary, the MLC leaves move along the International Electrotechnical Commission (IEC) x-direction, dynamically adapting in real-time to ensure the radiation beam continuously follows the time-dependent tumor position [32].

MR-Linacs enable beam delivery with greater conformality compared to conventional IGRT, and the research community has demonstrated a rapidly increasing interest in the role of MRI in radiotherapy as well as its applications in motion management. For a detailed review of MRgRT, including its current status and future roadmap, please refer to [16, 79].

1.4 Latency and motion-related imaging errors in MRgRT

1.4.1 Characterization and impacts

Despite the aforementioned advantages of MRgRT, its status as a relatively new technology indicates ongoing potential for optimization. Given the sensitivity of TCP and NTCP to the prescribed dose, achieving higher precision in dose delivery remains essential. Consequently, the gating and tracking performance of MR-Linacs has become a primary focus.

Latency serves as a key indicator for evaluating the accuracy of beam gating and MLC tracking. Gating latency refers to the delay between the target's status change and the corresponding beam resumption or cessation. It is typically divided into beam-on latency, the delay between the target entering the gating window and the initiation of treatment, and, more importantly, beam-off latency, the delay between the target exiting the window and the beam being switched off. Kim et al. [80] measured the latency for the ViewRay MRIdian MR-Linac, reporting a largest measured beam-off latency of 302 $\pm$ 20 ms with 8-FPS cine MRI, with average values ranging from 128–243 ms for 4 Hz Cartesian acquisition and 47–302 ms for 8 Hz radial acquisition. The end-to-end MLC-tracking latency can be defined as the delay between a moving target reaching a specific position and the center of the MLC leaves following that target arriving at the same position. Glitzner et al. [32] conducted a technical study on the Elekta Unity MR-Linac and reported MLC-tracking latencies of 347.45 ms at 4 Hz imaging and 204 ms at 8 Hz. Liu et al. [33] from the Australian MR-Linac project measured a time delay of 328 ± 44 ms in the MLC beam-repositioning response. These latencies lead to gating and tracking errors, which are especially critical when rapid target motion occurs due to respiration or cardiac activity. As they represent a root cause of dose coverage loss, such latencies should be minimized as much as possible [32, 86, 87].

The sources of overall loop latency in the gating or tracking workflow with an MR-Linac can generally be categorized according to the stages of the process: imaging latency, image processing (e.g., contouring or target tracking algorithms), and machine control (e.g., MLC leaf repositioning and beam triggering latency). Imaging latency is defined as the delay between the occurrence of a physical change and its representation in the reconstructed image [88]. This latency can be further broken down into components associated with data acquisition, as well as the time required for non-zero data transfer and reconstruction. According to the above-mentioned latency experiments reported in the literature, MR imaging latency has been identified as the largest contributor to the total end-to-end latency in real-time MRI-based adaptive radiotherapy. Liu et al. [33] measured it as 194 ± 43 ms, with 69 ± 42 ms attributed to reconstruction and data transfer, and 125 ± 5 ms due to acquisition; Glitzner et al. [32] concluded that MLC delays are negligible, as the latency and geometric tracking errors induced by MR imaging exceed the MLC-related errors by several factors. They further confirmed that optimizing the MRI acquisition process offers the greatest potential for advancing real-time motion management in MRgRT. Additionally, with continued advancements in reconstruction algorithms and computing capabilities, the relative impact of data transfer and reconstruction—already minor contributors—can be further diminished. Although techniques such as partial Fourier and undersampling have been adopted to accelerate image acquisition [89], the latency contribution from this stage, being the largest component, remains a central concern.

Imaging latency associated with the acquisition process can be viewed as a manifestation of motion-related imaging errors. Unlike static imaging, real-time MR imaging employed for motion monitoring (cine-MRI in contemporary MR-Linacs) captures dynamic anatomical structures. Given that the acquisition time for a single cine-MR frame is comparable to the timescale of physiological motion, the finally acquired k-space incorporates signals of the target at varying positions. This manifests in the image domain as motion-induced errors, which differ in origin from static imaging errors, such as blurring from limited spatial resolution or artifacts due to undersampling.

Motion-related imaging errors in cine-MR frames can be approached from two perspectives: image blurring and target positioning errors. The extent of motion-induced image blurring or artifacts is generally discernible and can be readily assessed by domain experts. In contrast, inaccuracies in target positioning reflect a lack of real-time responsiveness in the imaging process, indicating that the apparent position or geometry of the object derived from the image lags behind its actual position or geometry at the end of acquisition, irrespective of image reconstruction. These errors correspond to the contribution of imaging latency related to the acquisition process. Compared to image blur, target positioning errors are more difficult to perceive or detect, making them prone to being overlooked [90]. Previous motion correction techniques for conventional MR systems [91–94] have primarily focused on mitigating image blur for diagnostic purposes. However, target positioning errors are particularly relevant in the context of MR-guidance for active motion management, where accurate and up-to-date localization of organ position and geometry is crucial.

Hereinafter, the physical motion of objects occurring within a single cine-MR frame acquisition is referred to as intra-frame motion in this work. Organ motion due to respiration has been extensively measured [20]. Published results indicate that fast anatomical variations within a single breathing cycle can be expected, particularly in a deep breathing mode, where diaphragm motion amplitudes of up to 101 mm along the superior-inferior (SI) direction have been observed [95]. Additionally, cardiac activity has been found to substantially contribute to rapid positional changes of lung tumors, mediastinal lymph nodes, or liver tumors [96–98]. Observations on lung tumor motion showed that the speed can reach up to 72.6 ± 22.5 mm/s [99]. Given the relatively long time span of the acquired k-space data points, effective intra-frame motion can be involved, leading to appreciable motion-related imaging errors. Moreover, to achieve potential heart dose reduction [100], real-time motion monitoring of the heart is required, imposing higher demands on dynamic imaging performance owing to the rapid anatomical deformation induced by cardiac activity.

1.4.2 Current solutions and mitigation strategies

To account for system latencies, motion prediction techniques [101–103] have been developed to forecast future organ positions based on previously observed motion states, with particular focus on respiratory motion. However, as stated in the AAPM TG76 report, "There are no general patterns of respiratory behavior that can be assumed for a particular patient prior to observation and treatment" [20]. The individual characteristics and natural variability of respiration present intrinsic challenges for motion prediction, especially in accurately capturing the turning points of the motion amplitude curve. Additionally, recent studies have demonstrated that predicting 2D geometries, such as tumor contours, is more challenging than predicting the 1D position of tumor centroids [104]. As a result, mitigating residual geometric tracking errors in imaging of large deformable targets remains a critical challenge.

Borman et al. [88] aimed to mitigate imaging latency associated with the acquisition process by altering the phase-encoding order in Cartesian readouts. However, the proposed effective *high-low* ordering scheme can introduce more significant eddy current artifacts, necessitating additional compensation strategies. Moreover, in *high-low* orderings, larger discrepancies in the higher frequency components (HFC) between the obtained motion-corrupted image and the ground-truth final-position image can arise, as these components are acquired earlier. This may still pose challenges to image quality, since HFCs carry valuable semantic information that can be leveraged by certain algorithms for contouring [105–107].

Radial trajectories are robust to a certain level of azimuthal undersampling, enabling sliding window reconstruction at nearly arbitrary frame rates. However, motion-related imaging errors are independent of the frame rate and instead correlate with the temporal span of spokes within the reconstruction window. Highly undersampled image reconstruction [108, 109] is a common technique for reducing imaging errors in radial cine-MR, involving image restoration through approaches such as artificial intelligence (AI) algorithms under the constraint of a narrower reconstruction window. However, this technique is limited to a certain acceleration factor, as higher factors demand more semantic reasoning about the input [110]. Borman et al. [88] applied a spatial-temporal (k-t) filter in golden-angle sequences, retrospectively downscaling the lower-frequency components (LFC) of previously acquired spokes while preserving HFCs. This approach reduced the imaging latency by approximately 50%, demonstrating room for further optimization.

1.5 Specific aims and thesis outline

Object motion-induced deterioration of image quality in radiography have been thoroughly studied and mathematically characterized using impulse responses, which are subsequently represented by the modulation transfer function (MTF) formalism to combine both spatial and temporal degradation of the imaging system [111, 112]. For example, uniform motion at a given velocity smears a point into a line, resulting in a box-function impulse response, in which case the system's spatial MTF needs to be multiplied by a sinc function. However, MR imaging is essentially different, as the raw signal is acquired in the Fourier domain. Therefore, dedicated efforts are required to address this dynamic imaging behavior and to mitigate residual tracking errors of MR-guidance, particularly in cases of fast breathing or for anatomical structures affected by the heartbeat.

The primary aim of this thesis is to investigate motion-related errors in real-time MR imaging and explore the feasibility of reducing these errors through deep learning-based intra-frame motion compensation techniques. Unlike previous motion correction methods, which primarily aim to restore motion-blurred images, the proposed compensation techniques also focus on addressing target positioning errors to enhance the real-time responsiveness of MRI. In contrast to motion prediction methods, the compensation approach aims to directly extract and derive the implicit real-time position from the given input, rather than forecasting based on prior knowledge. For radial sampling, as opposed to high-undersampling techniques, the compensation methods are expected to retain an adequate reconstruction window width, thereby preserving more semantic information for high image fidelity, based on the hypothesis that earlier-acquired spokes still contribute to the estimation. In line with these objectives, this thesis is structured as follows:

Chapter 2 provides the basic physical and technical background of MRI, linear accelerators, and the MR-Linac system, offering a general understanding of MR imaging principles and MRgRT.

Chapter 3 presents the development and validation of a simulation framework for motion-dependent MRI sampling, designed to support a fundamental understanding of motion-related imaging errors. The simulation reveals that frequency-domain information corresponding to each temporal position is spatially and temporally encoded within the k-space of motion-corrupted images. Based on this insight, an inverse problem is formulated to recover the real-time final-position image, corresponding to the end of the frame acquisition, directly from the motion-corrupted data.

Over recent years, deep-learning algorithms have played an increasingly important role in MRI or MRgRT across various applications, including motion correction [93, 94], image segmentation [113, 114], synthetic CT generation [115] and online treatment planning [116]. Numerous studies have highlighted the transformative potential of deep learning in these areas [117]. By learning a mapping function from the input space to the output space, neural networks emerge as a prominent solution for addressing the inverse problem central to this thesis.

A key determinant of the network's success lies in the creation of a suitable training dataset. Accordingly, Chapter 3 details the methodology for generating intra-frame motion datasets based on digital phantoms, with the simulation procedure serving as a generator for labeled training pairs. The process involves the development of 4D MRI digital anthropomorphic phantoms, followed by the introduction of an intra-frame motion model and a motion pattern perturbation scheme.

Given the distinct characteristics of Cartesian and radial k-space readout trajectories, each sampling pattern requires a network architecture uniquely designed to accommodate its specific needs. Chapter 4 and Chapter 5 conduct a proof-of-concept study on reducing imaging errors through the implementation of deep learning-based intra-frame motion compensation techniques, with a focus on Cartesian and radial cine-MRI, respectively.

Chapter 4 begins by discussing the rationale for selecting a UNet architecture tailored to Cartesian sampling. The network's performance is subsequently evaluated on both fully sampled and undersampled Cartesian datasets. Furthermore, to enhance interpretability, saliency maps are analyzed in both the image and Fourier domains, highlighting the regions in the motion-corrupted image or k-space that contribute most significantly to the model's inference.

In Chapter 5, a novel intra-frame motion compensation network, TransSin-UNet, is introduced to address the challenges of radial k-space sampling without compromising the reconstruction window width. The network operates in both the projection and spatial domains, where long-range spatial-temporal dependencies among the sinogram representations of the radial spokes are modeled by a transformer encoder subnetwork, followed by a UNet subnetwork for pixel-level fine-tuning in the spatial domain. The network is then trained and extensively evaluated across datasets characterized by varying azimuthal radial profile increments.

The thesis is concluded by Chapter 6, where the main research findings are summarized, and future research perspectives are outlined.

Chapter 2

PHYSICAL AND TECHNICAL BACKGROUND

2.1 Nuclear magnetic resonance

2.1.1 Spin angular momentum and magnetic moment

MRI is based on nuclear magnetic resonance (NMR). The magnetism of the nucleus originates from its magnetic moment, which in turn arises from the spin angular momentum of the nucleus.

Spin is the quantum mechanical property of elementary and composite particles that is associated with their intrinsic angular momentum. The total nuclear angular momentum J is quantized and can be written as:

$$J = \hbar\sqrt{I(I+1)} \tag{2.1}$$

where $\hbar$ is the reduced Planck constant, given by $\hbar = h/2\pi$; I refers to the spin quantum number, $I = n/2$, where n can be any non-negative integer. When the atomic mass number A is odd, I takes half-integer values. For example, for ^{1}H (proton), ^{13}C, ^{15}N, ^{19}F, etc., $I = 1/2$; for ^{7}Li, ^{9}Be, ^{23}Na, etc., $I = 3/2$. When the mass number A is even and the atomic number Z is odd, I takes integer values. For example, for ^{2}H, ^{14}N, $I = 1$. When both the mass number and atomic number are even, $I = 0$, as seen in ^{4}He, ^{12}C, which do not have spin angular momentum.

The magnetic moment is directly related to the spin angular momentum vector as follows:

$$\boldsymbol{\mu} = \gamma \mathbf{J} \tag{2.2}$$

The proportionality constant γ in Eq. 2.2 is called the gyromagnetic ratio and depends on the particle or nucleus. For the proton, its value is found to be:

$$\gamma = 2.675 \times 10^{8} rad/s/\mathrm{T} \tag{2.3}$$

While MRI can theoretically be performed with all nuclei with a non-zero spin, proton NMR is mainly exploited in clinical routine, due to the high concentration of hydrogen atoms in the human body (88 mol/L) [118], as well as the relatively large gyromagnetic ratio, which provides a higher detectable signal.

2.1.2 Macroscopic magnetization

In the absence of an external magnetic field, spin orientations are randomly distributed. However, when placed in a strong external magnetic field $\mathbf{B_0}$, oriented along the z-direction, the spin experiences a torque. Constrained by the laws of quantum mechanics, the magnetic moment cannot align exactly with the $\mathbf{B_0}$ direction, but instead maintains a specific angle, subjecting it to a constant torque. This torque induces the magnetic moment to precess counter-clockwise around the main field B_0 with a constant angular frequency, known as the Larmor frequency ω_0, computed as:

$$\boldsymbol{\omega_0} = \gamma \mathbf{B_0} \tag{2.4}$$

The angular momentum component along the z-axis, J_z, is quantized:

$$J_z = \hbar I_z = \hbar m; \quad m = -I, -I+1, ..., I-1, I \tag{2.5}$$

Here, m is the magnetic quantum number, which can take on $2I+1$ possible values, corresponding to $2I+1$ magnetic energy levels E_m:

$$E_m = -\boldsymbol{\mu} \cdot \mathbf{B_0} = -\mu_z B_0 = -\gamma \hbar I_z B_0 = -\gamma \hbar m B_0 \tag{2.6}$$

This phenomenon is known as the Zeeman effect. Consequently, for a proton with $I = 1/2$, there are only two possible states, defined by $m = \pm 1/2$. The lower-energy state is almost aligned with the main field, referred to as spin-up or parallel. The other higher-energy state, known as spin-down or anti-parallel, is aligned almost opposite to the external field. The energy difference between these two states is calculated as:

$$\Delta E = \gamma \hbar B_0 \tag{2.7}$$

In thermal equilibrium, the number of particles P_m in each state follows the Boltzmann distribution:

$$P_m \propto exp(-E_m/kT) \tag{2.8}$$

where k is the Boltzmann constant, $k = 1.380649 \times 10^{-23}$J/K; and T is the thermodynamic temperature. Therefore, the net magnetization M_0 in a given volume V sample containing N nuclear spins is given by:

$$M_0 = \frac{N}{V}\langle\mu_z\rangle = \frac{N}{V}\cdot\gamma\hbar\,\frac{\sum\limits_{m=-I}^{I} m\exp\left(\frac{\gamma B_0 m\hbar}{kT}\right)}{\sum\limits_{m=-I}^{I}\exp\left(\frac{\gamma B_0 m\hbar}{kT}\right)} \qquad [M_0] = \text{J T}^{-1}\text{m}^{-3} \tag{2.9}$$

where $\langle\mu_z\rangle$ is the expectation value of z-component of the magnetic moment. Since $\gamma\hbar B_0 \ll kT$ at body temperature and clinical field strengths, it is appropriate to perform a Taylor series expansion of the Boltzmann exponential function and apply the first-order approximation. Note that $\sum\limits_{m=-I}^{I} m = 0$, $\sum\limits_{m=-I}^{I} m^2 = I(I+1)(2I+1)/3$ and $\sum\limits_{m=-I}^{I} 1 = 2I+1$. Substituting these expressions, Eq. 2.9 can thus be simplified as:

$$M_0 = \frac{N}{V}\frac{\gamma^2\hbar^2 I(I+1)}{3kT}B_0 \tag{2.10}$$

It can be found that the lower-energy state is slightly favored, causing the direction of the net magnetization to align exactly in parallel with the main field B_0.

For the proton, replacing N/V with the proton density ρ, the net magnetization M_0 becomes:

$$M_0 = \frac{\rho\gamma^2\hbar^2 B_0}{4kT} \tag{2.11}$$

2.1.3 Resonance transition

To measure a signal, a radio frequency (RF) pulse is applied to induce transitions between the two states, tipping M_0 to have a component in the x-y transverse plane. For a transition to occur, the RF energy must match the difference between the two energy states, $\Delta E = \hbar\omega_0$. Therefore, the oscillation frequency of the RF pulse should satisfy the resonance excitation condition, which means it must be equal to the Larmor frequency, ω_0. For an MR-Linac with a magnetic field strength of 0.35 T, the resonance frequency, $f = \omega_0/2\pi = 14.90$ MHz, while for $B_0 = 1.5$ T, $f = 63.87$ MHz. These frequencies fall within the radio wave range and are far lower than the frequencies of ionizing radiation.

The RF pulse is produced either linearly or circularly polarized and creates a fixed magnetic field $\mathbf{B_1}$ along the x-axis in the ω_{rot} rotating frame, where ω_{rot} denotes the frame's rotation frequency. The temporal evolution of the magnetization $\mathbf{M}$ in the rotating frame is given by:

$$\left(\frac{d\mathbf{M}}{dt}\right)_{rot} = \gamma\mathbf{M} \times \left(\left(\mathbf{B_0} - \frac{\omega_{rot}}{\gamma}\right)\hat{e}_z + \mathbf{B_1}\hat{e}_x\right) = \gamma\mathbf{M} \times \mathbf{B_1}\hat{e}_x \tag{2.12}$$

provided by $\omega_{rot} = \omega_0$. In this rotating frame, the magnetization $\mathbf{M}$ rotates about the x-axis until the RF pulse is switched off; in the fixed laboratory frame, this corresponds to a spiraling motion away from the z-axis.

The flip angle α, achieved at the end of the RF pulse with a duration t_p, is given by:

$$\alpha = \gamma B_1 t_p \tag{2.13}$$

When $\alpha = 90°$, the RF pulse is referred to as a $90°$-pulse, which rotates the magnetization into the transverse plane.

After the RF pulse application, the magnetization vector $\mathbf{M}$ precesses around the z-axis with the Larmor frequency ω_0, with a longitudinal component $M_z = M\cos\alpha$ and a transverse component $M_{xy} = M\sin\alpha$. The transverse component induces a voltage in the receiver coil. The signal amplitude decays exponentially to zero within only a few milliseconds as the protons rapidly dephase with respect to each other. This signal is known as the free induction decay (FID) signal (Section 2.3).

2.2 Relaxation

For MR imaging, it is essential to differentiate tissues, ensuring that identical tissues produce the same signal values, while distinct tissues yield different values. This can be achieved by exploiting tissue-specific parameters, including the proton density ρ, as well as the longitudinal and transverse relaxation times, T_1 and T_2.

Having excited the protons, the magnetization begins to relax back to the equilibrium position as soon as the RF pulse is switched off. The temporal evolution of the magnetization during excitation and relaxation can be described by the Bloch equations:

$$\frac{d\mathbf{M}}{dt} = \gamma\mathbf{M} \times \mathbf{B} = \gamma \begin{pmatrix} M_y B_z - M_z B_y \\ M_z B_x - M_x B_z \\ M_x B_y - M_y B_x \end{pmatrix} \tag{2.14}$$

Therefore, during relaxation, the variations of longitudinal and transverse components can be written as:

$$\begin{aligned}\frac{\mathrm{d}M_z(t)}{\mathrm{d}t} &= \frac{M_0 - M_z(t)}{\mathrm{T}_1} \\ \frac{\mathrm{d}M_{xy}(t)}{\mathrm{d}t} &= -\frac{M_{xy}(t)}{\mathrm{T}_2}\end{aligned} \tag{2.15}$$

The solution to Eq. 2.15 in the rotating frame is:

$$\begin{aligned}M_z(t) &= M_0 - (M_0 - M_z(0)) \cdot \exp\left(-\frac{t}{\mathrm{T}_1}\right) \\ M_{xy}(t) &= M_{xy}(0) \cdot \exp\left(-\frac{t}{\mathrm{T}_2}\right)\end{aligned} \tag{2.16}$$

M_{xy} and M_z correspond to different relaxation features. The recovery of the magnetization along the z-axis is referred to as spin-lattice relaxation, or T_1 recovery, which results from the interaction of protons with the surrounding environment (the lattice). In this process, the spin system loses excess energy from the RF pulse and returns to the thermal equilibrium state. Let $t = \mathrm{T}_1$ and $M_z(0) = 0$ in the first equation of Eq. 2.16, it can be found that T_1 is the time required for the longitudinal magnetization component to recover from 0 to 63% of its equilibrium value.

The exponential decay of magnetization in the transverse plane is referred to as spin-spin relaxation, or T_2 decay, and arises from the dephasing of spins after the RF pulse. This dephasing occurs due to slight variations in the precession frequencies of spins, which are induced by random interactions between spins that generate internal field inhomogeneity. When two protons are in close proximity, the magnetic moment of one proton either enhances or diminishes the local magnetic field experienced by the other. There is no net energy loss of the spin system in this process. Let $t = \mathrm{T}_2$ in the second equation of Eq. 2.16, it can be found that T_2 is the time required for the transverse magnetization component to decay to 37% of its initial value. In human tissues, T_2 is always shorter than T_1.

In practical MR imaging, the transverse magnetization decays at a rate of $1/\mathrm{T}_2^*$, which is significantly faster than $1/\mathrm{T}_2$. In addition to internal field inhomogeneities, external field inhomogeneities also contribute to dephasing. This effect is characterized by a distinct relaxation time T_2'. Unlike T_2 decay, T_2' decay can be recovered by creating an echo. The overall transverse relaxation time, T_2^*, is expressed in terms of T_2 and T_2':

$$\frac{1}{\mathrm{T}_2^*} = \frac{1}{\mathrm{T}_2} + \frac{1}{\mathrm{T}_2'} \tag{2.17}$$

and $M_{xy}(t)$ in Eq. 2.15 becomes:

$$M_{xy}(t) = M_{xy}(0) \cdot \exp\left(-\frac{t}{\mathrm{T}_2^*}\right) \tag{2.18}$$

The proton density and relaxation times can lead to variations in the transverse magnetization across different tissues, even when their elemental composition is similar, which explains the high soft tissue contrast of MRI.

2.3 Free induction decay and echo creation

Corresponding to the representation in the rotating frame described by Eq. 2.18, the transverse magnetization in the fixed laboratory frame can be expressed in complex form as:

$$M_{xy}(t) = M_{xy}(0) \cdot \exp\left(j\omega_0 t - \frac{t}{\mathrm{T}_2^*}\right) \tag{2.19}$$

which generates a measurable time-dependent FID signal, $S(t)$, in the receiver coil:

$$S(t) = \int M_{xy}(\boldsymbol{r}, t)\; d^3\boldsymbol{r} \tag{2.20}$$

where $\boldsymbol{r} = (x, y, z)$ represents the position vector in three-dimensional space. The signal exhibits a damped oscillatory behavior, characterized by the Larmor frequency ω_0 and an exponentially decreasing amplitude.

As mentioned earlier, spin-spin relaxation caused by external field inhomogeneities, described by T_2', is a reversible mechanism that can be compensated by generating an echo. In MR imaging, there are two types of echoes: spin echo (SE) and gradient echo (GE).

For SE, after the excitation RF pulse, a $180°$-pulse is applied at time $T_E/2$, causing the spins to rotate $180°$ about the x-axis and flip to the mirror position. After an additional $T_E/2$, the transverse components of the spins rephase again, which reverses the T_2' decay process. At this moment, the resulting echo peak is :

$$M_{SE} = M_{xy}(0) \cdot \exp\left(-\frac{T_E}{\mathrm{T}_2}\right) \tag{2.21}$$

After reaching its peak, the echo undergoes further decay according to T_2^*, similar to the FID signal, as expressed by:

$$M_{\text{echo}}(t) = \left(M_{xy}(0) \cdot \exp\left(-\frac{T_E}{\mathrm{T}_2}\right)\right) \cdot \exp\left(-\frac{t - T_E}{\mathrm{T}_2^*}\right) \tag{2.22}$$

In GE, a magnetic gradient field is applied immediately after the excitation RF pulse, superimposed on the main field. This introduces spatial variations in the Larmor frequency, causing the transverse magnetization to dephase more rapidly than the FID signal. After a defined period, this dephasing is counteracted by applying an inverted gradient field of equal strength but opposite polarity. At echo time T_E, the spins rephase, forming a gradient echo, with the echo peak given by:

$$M_{GE} = M_{xy}(0) \cdot \exp\left(-\frac{T_E}{\mathrm{T}_2^*}\right) \tag{2.23}$$

The MR imaging sequences (Section 2.6) commonly employed in clinical practice are derived from the basic SE and GE pulse sequences.

2.4 Spatial Encoding

This section uses 2D Cartesian k-space sampling (Section 2.5), the most common and conventional MRI signal acquisition method, as an example to introduce the principles of MRI spatial encoding.

For MR imaging, it is essential to represent magnetization using gray-scale values as a function of spatial coordinates, i.e., $M_{xy}(x, y, z)$. The signal in Eq. 2.20 corresponds to the integral of transverse magnetization over all excited regions within the volume. To obtain the spatial distribution of M_{xy} and reconstruct an MR image, spatial encoding is implemented by applying magnetic gradient fields that are superimposed onto the main magnetic field $\mathbf{B_0}$. These gradient fields are oriented parallel to $\mathbf{B_0}$ and vary linearly in strength along the x-, y-, and z-axis, with slopes of G_x, G_y, and G_z, respectively. Each gradient enables a distinct form of spatial encoding: slice selection, phase encoding and frequency encoding. All spatial encoding relies on the fact that the Larmor frequency is proportional to the magnetic field strength (see Eq. 2.4).

2.4.1 Slice selection

The slice-selective gradient, such as G_z, confines the excited region to a slice of finite thickness. In the presence of G_z, the Larmor frequency ω at position z is:

$$\omega = \gamma\left(B_0 + G_z \cdot z\right) \tag{2.24}$$

When an RF pulse with a bandwidth BW is applied during the presence of G_z, a specific slice is excited with a thickness Δz, given by:

$$\Delta z = \frac{BW \cdot 2\pi}{\gamma G_z} \tag{2.25}$$

The location of the excited slice can be controlled by adjusting the center frequency of the RF pulse or modifying the amplitude offset of G_z.

2.4.2 Phase and frequency encoding

After the slice selection, the transverse magnetization at each location in the x-y plane is obtained by repeatedly modulating the integrated signal, introducing spatially varying phases to the spins across different positions. This approach is inspired by the two-dimensional discrete Fourier transform (DFT):

$$S(k_x, k_y) = \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} M_{xy}(x, y) \cdot \exp\left[-j\left(\frac{2\pi}{N_x}k_x x + \frac{2\pi}{N_y}k_y y\right)\right] \tag{2.26}$$

where k_x and k_y are the spatial frequencies along the x and y directions, respectively; and $\frac{2\pi}{N_x}k_x x$ and $\frac{2\pi}{N_y}k_y y$ correspond to the phase shifts, which can be further expressed as:

$$\begin{aligned} \frac{2\pi}{N_x}k_x x = \varphi_x(x) = \gamma G_x x \Delta t_x \\ \frac{2\pi}{N_y}k_y y = \varphi_y(y) = \gamma G_y y \Delta t_y \end{aligned} \tag{2.27}$$

Thus, the variables k_x and $k_y$can be manipulated either by fixing G_x or G_y and varying the respective gradient durations Δt_x or Δt_y, or by keeping Δt_x or Δt_y constant while varying G_x or G_y. The specific configuration depends on the assignment of the frequency and phase encoding directions to the respective axes.

Assume a configuration where frequency encoding is applied along the x-axis and phase encoding along the y-axis. Accordingly, along the y-axis, phase encoding is performed by activating the gradient field G_y for a duration of Δt_y. During this period, spins at different y-positions precess at slightly different Larmor frequencies. Once the gradient field is switched off, the transverse magnetization vectors at different y-positions return to the same frequency value but have accumulated distinct phase shifts φ_y. Along the x-axis, frequency encoding is performed by activating the gradient field G_x during the signal readout for a duration of Δt_x. Spins at different x-positions precess at slightly different Larmor frequencies, resulting in transverse magnetization vectors with position-dependent phase shifts φ_x.

To reconstruct an MRI image, signals corresponding to different k_x and k_y values must be acquired. In this encoding configuration, multiple k_x values can be sampled within a single signal readout evolution by adjusting the sampling time to obtain distinct Δt_x values. In contrast, different k_y values are controlled by varying G_y, which requires multiple signal readout iterations. Each iteration employs a unique G_y to sample a specific k_y value. Consequently, the temporal scale associated with the k_x

axis is negligible compared to that of k_y. Given a repetition time T_R, which defines the duration of each signal readout iteration, the total MRI slice acquisition time T_{acq} is determined by the number of pixels N_p along the phase encoding direction:

$$T_{acq} = N_p \cdot T_R \tag{2.28}$$

2.5 *k*-space

The k-space matrix is the repository for discretized spatial frequency signals acquired during the evolution and decay of the echo, which is equivalent to the Fourier space of spatial frequencies. After shifting the direct current (DC) component to the center, the frequency axes become symmetric about the center of k-space, spanning from $-k_{max}$ to $+k_{max}$.

The strength and direction of the magnetic field gradient can be rapidly modulated over time, allowing for different possible trajectories to acquire the k-space, such as Cartesian, spiral, and radial. Fig. 2.1 illustrates the Cartesian and radial sampling patterns, which are commonly used in conventional 2D cine-MRI. In a Cartesian trajectory, sampling points are arranged in a uniform grid along the Cartesian coordinate system, with each readout iteration filling a horizontal line of k-space, corresponding to one phase encoding step. In contrast, radial sampling distributes points along radial profiles, known as spokes, that pass through the center of k-space. Upon completion of the current readout iteration, the gradient field shifts the acquisition to the next Cartesian line or radial spoke. Compared to Cartesian trajectories, radial sampling is less sensitive to motion artifacts and offers greater robustness against a certain level of azimuthal undersampling, thereby enhancing temporal resolution. Additionally, it enables sliding window reconstruction at a nearly arbitrary frame rate. These properties make the radial readout trajectory a preferred choice for imaging dynamic physiological processes, but it requires relatively more complex reconstruction techniques (see Section 2.7.1).

Since the spatial and frequency domains are related by the Fourier transform (see Eq. 2.26), each data point of k-space represents the signal amplitude contributed by all MR image voxels corresponding to that specific spatial frequency components. Fig. 2.2 illustrates an 11×11 matrix that visualizes the image patterns associated with signals at different k-space spatial frequency coordinates (k_x, k_y), where the real part of the signal values is considered. It is evident that as the location moves from the center to the periphery of k-space, the number of line pairs per unit distance increases along the k_x- and k_y-axes.

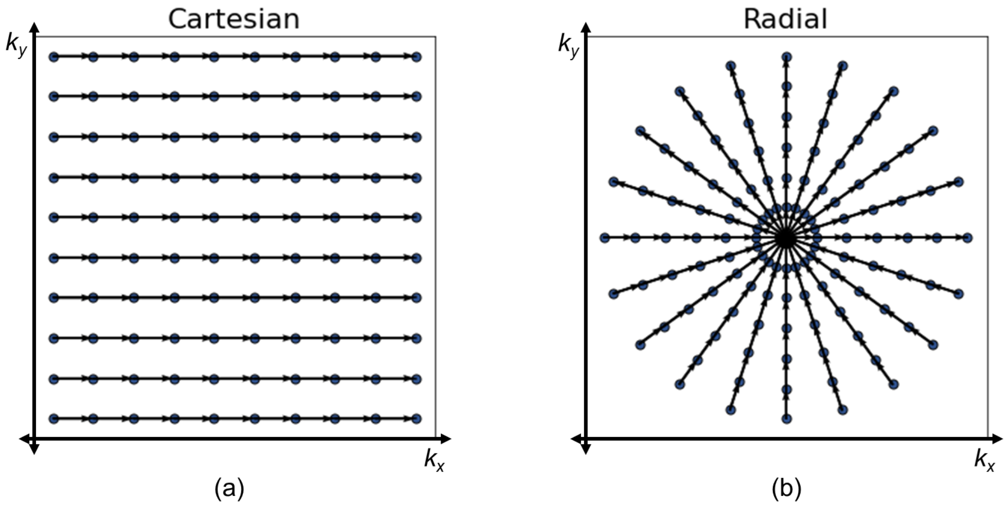

Figure 2.1: Illustration of (a) Cartesian and (b) radial k-space sampling trajectories. Each line corresponds to a signal readout evolution, where dots indicate acquired samples and arrows denote the sampling direction. In the Cartesian trajectory, each readout iteration fills a horizontal line of k-space, whereas in the radial trajectory, it fills a radial profile passing through the center, known as a spoke. Upon completion of the current iteration, the gradient shifts the trajectory to the next line.

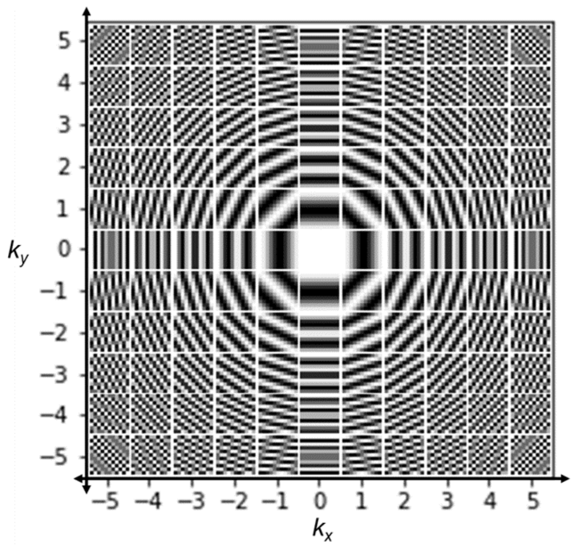

Figure 2.2: Image patterns (real part) associated with signals at different k-space spatial-frequency coordinates. From the center to the periphery, the number of line pairs per unit distance increases along the k_x- and k_y-axes.

2.6 Imaging sequences

An MR imaging sequence is a prescribed arrangement of RF excitation pulses and magnetic field gradient applications, designed to achieve spatial encoding and signal readout for producing images with specific characteristics.

Fig. 2.3 presents a schematic of imaging sequences for 2D Cartesian and radial acquisitions. During the application of the RF pulse, the slice selection gradient G_{SS} is activated, followed by a free precession period. For a Cartesian acquisition, during this period, a k_y-value is encoded by the phase-encoding gradient G_{PE}. The transverse magnetization is dephased and rephased by the read-out gradient G_{RO} with opposite polarities, leading to the formation of a gradient echo after the echo time T_E. In contrast, for a radial acquisition, gradient fields along the x and y directions, G_x and G_y, are applied simultaneously, following:

$$\begin{aligned} G_x &= G_{\mathrm{RO}} \cdot \cos\theta \\ G_y &= G_{\mathrm{RO}} \cdot \sin\theta \end{aligned} \tag{2.29}$$

where θ represents the angle between the current radial spoke and the x-axis. The magnetization is dephased by a spoiler gradient before the pulse sequence is repeated after the repetition time T_R. In the Cartesian acquisition, this repetition occurs with a different G_{PE} for the next phase-encoding step, while in the radial acquisition, it proceeds with a different θ value for the next spoke sampling.

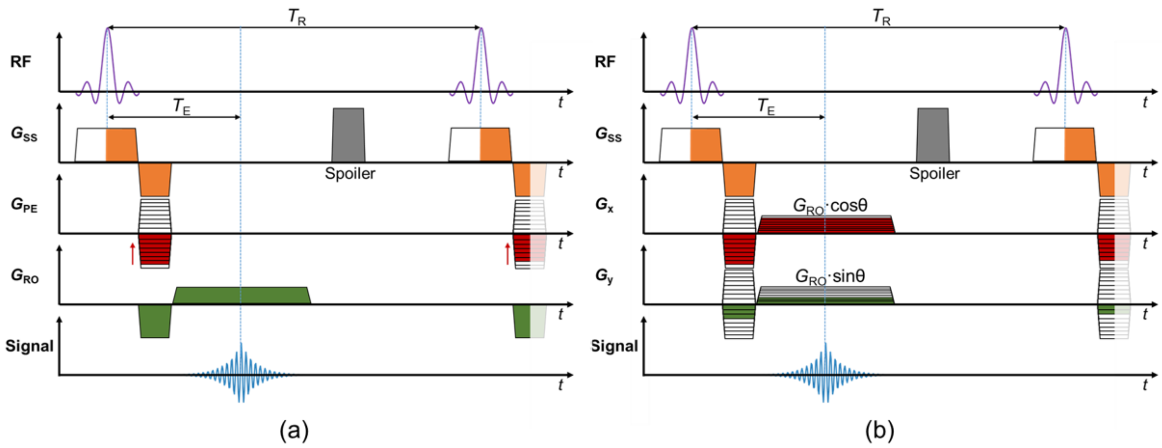

Figure 2.3: Illustrative representation of imaging sequences. (a) Gradient echo sequence for Cartesian acquisition. (b) Imaging sequence for radial acquisition. RF: radio frequency pulse. SS: slice selection. PE: phase-encoding. RO: read-out (frequency-encoding). TE: echo time. TR: repetition time.

The strength of the MRI signal depends on both sequence-specific parameters, such as T_R, T_E, the flip angle α, the RF bandwidth, the gradient strengths, the inversion time T_I (used in inversion recovery sequences [119]), as well as tissue-specific parameters, including the proton density ρ and T_1 and T_2 relaxation times. By designing the imaging sequence and varying sequence-specific parameters, different k-space sampling schemes and multiple MRI modalities with distinct tissue contrast can be achieved, enabling a range of clinical applications, such as the acquisition of T_1-weighted, T_2-weighted, or proton density weighted images, or the suppression of signals from specific tissues using additional preparation pulses.

2.7 Image reconstruction and acceleration techniques

2.7.1 Image reconstruction

An MR image reflects the distribution of transverse magnetization M_{xy} and as indicated by Eq. 2.26, image reconstruction for Cartesian sampling patterns can be performed using a simple inverse Fourier transform (IFT) of the acquired k-space data, with its discrete form given by:

$$M_{xy}(x, y) = \sum_{k_x=0}^{N_x-1} \sum_{k_y=0}^{N_y-1} S(k_x, k_y) \cdot \exp\left[j\left(\frac{2\pi}{N_x}k_x x + \frac{2\pi}{N_y}k_y y\right)\right] \tag{2.30}$$

The fast Fourier transform (FFT) along with its inverse (IFFT) constitutes an efficient algorithm for computing the discrete Fourier transform and its inverse. By recursively breaking down the DFT into smaller DFTs and exploiting symmetries in the transform, the FFT significantly reduces the computational complexity from $O(N^2)$ (which results from directly applying the DFT definition) to $O(N \log N)$, where N is the number of data points, thereby minimizing redundant calculations.

For non-Cartesian sampling patterns, such as radial and spiral, reconstruction methods based on inverse non-uniform fast Fourier transform (NUFFT) [120, 121] are required. The basic idea of NUFFT is to interpolate Cartesian samples from adjacent non-uniformly acquired data points, as illustrated in Fig. 2.4. Gridding is a widely adopted approach that first convolves the acquired data with a kernel and then resamples it onto an oversampled uniform grid. Finer sampling in spatial frequency helps shift the sampling replicas further outward, forming a transition band that mitigates aliasing artifacts. The oversampling factor is typically chosen between 1.25 and 2 [122]. Among various kernel choices, the shift-invariant Kaiser-Bessel kernel is widely preferred in the MR community due to its effectiveness in minimizing aliasing errors [123].

The apodization effect introduced by the gridding kernel can be corrected by dividing the reconstructed data by the inverse transform of the kernel. Density correction is necessary because the sample density typically varies in non-Cartesian acquisitions. The density can be estimated based on analytical or numerical models, such as assigning an area to each sample [122], which can then be used as the density correction factor (DCF) to scale the sample value.

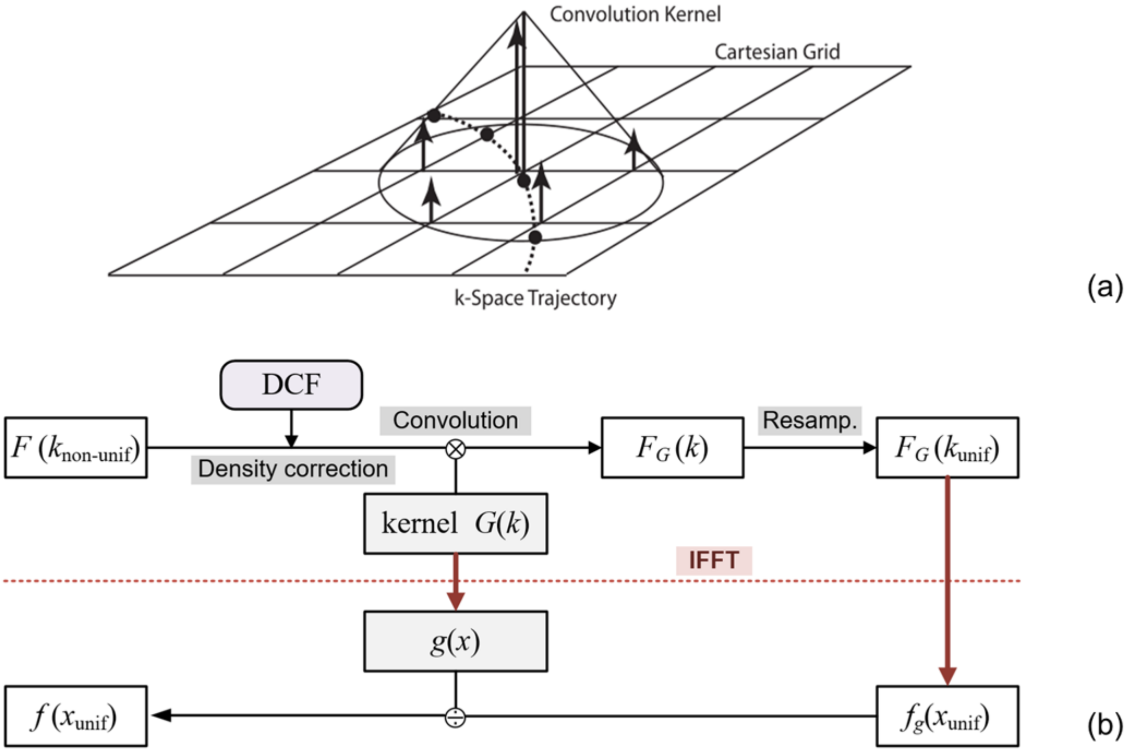

Figure 2.4: Illustration of the NUFFT reconstruction method. (a) Gridding kernel convolution and resampling (subfigure adapted from [122]). (b) Flowchart of the NUFFT process. "non-unif" denotes non-uniformly spaced samples, while "unif" denotes uniformly spaced samples.

2.7.2 Image acceleration

Accelerating data acquisition has been a long-standing goal in MRI. A straightforward approach is to collect fewer k-space samples than required. Nonetheless, k-space undersampling violates the Nyquist criterion and results in reconstruction artifacts. Over the past few decades, various techniques have been developed to reconstruct images of acceptable quality from undersampled data by exploiting intrinsic redundancies in MR images to recover the missing information [124].

Most clinically employed MRI protocols rely on partial Fourier and parallel imaging techniques for acceleration. The partial Fourier technique [125] takes advantage of the conjugate symmetry in the Fourier domain for real-valued images, where only a portion (down to 50% and typically 75%, referred to as the partial Fourier factor) of k-space is acquired instead of collecting the entire k-space. The missing data are then retrospectively inferred during the reconstruction. Parallel imaging approaches utilize multiple receiver coils in a phased-array setup, which assist with the spatial localization of the MR signal based on the known placement and local sensitivity of the different elements in the coils. This additional spatial information can be exploited to mathematically reconstruct the object of interest from undersampled k-space data. Reconstruction methods for parallel MRI with acceleration are mainly classified into two categories: image-domain-based and frequency-domain-based reconstruction methods. An example of the former class is the sensitivity encoding (SENSE) [126] method, which first reconstructs coil-specific images with IFFT and then unfolds aliased images based on the spatial sensitivity maps of the coils. The latter categories includes methods such as the generalized autocalibrating partially parallel acquisition (GRAPPA) [127], which first recovers the missing data in k-space based on, for example, auto-calibration signal (ACS) lines fitting, and then reconstructs the image using IFFT. By exploiting redundancies in the image domain or in k-space, partial Fourier and parallel imaging techniques typically achieve acceleration factors of 2–4$\times$ for most applications [124].

Compressed sensing (CS) [128, 129] has emerged as another powerful acceleration technique by leveraging the sparsity of MR images in a transform domain (e.g., wavelets). It enables higher acceleration factors and high fidelity image recovery from sparsely undersampled k-space data. However, its reliance on iterative reconstruction imposes substantial computational demands, restricting its applicability for real-time imaging.

Beyond these methods, deep learning has demonstrated significant potential in undersampled MRI reconstruction. In particular, data-driven neural networks learn priors from pre-constructed training pairs of undersampled data and their corresponding fully sampled counterparts, which serve as ground truth. This is achieved by iteratively updating the model parameters using an optimizer to minimize a predefined loss function that quantifies the discrepancy between the network output and the ground truth. The network can operate in the spatial domain, where it takes undersampled images with aliasing artifacts as input. In this case, architectures well-suited for image feature extraction, such as convolutional neural networks (CNN), have been widely explored [130]. Alternatively, the network can function in the Fourier domain by directly filling the undersampled k-space, where architectures designed to learn the inductive biases of frequency spectrum are actively being investigated [131].

Image acceleration involves a trade-off between image fidelity and acquisition time. In general, reducing the amount of available data increases uncertainty, thereby imposing a fundamental limit on the acceleration factor. Consequently, despite the availability of these acceleration techniques, the inherent nature of MRI imaging results in persistent imaging latencies, which manifest as motion-related imaging errors. Chapter 4 will present a case study where the proposed approach integrates undersampling-based acceleration with simultaneous motion-related imaging error reduction.

2.8 Main technical components

2.8.1 MR scanner components

Fig. 2.5 shows a schematic representation of the main technical components of an MR scanner. The static magnetic field B_0 is generated by the MR magnet. A superconducting magnet is one type that establishes an electric current in a loop of superconducting coils at temperatures approaching absolute zero (-273.16°C, 0 K), typically cooled with liquid helium at 4 K. The magnets are usually cylindrical in shape, with the patient placed inside the bore. Magnetic field gradients, G_x, G_y and G_z, used for spatial encoding along the three directions, are generated by gradient coils mounted inside the bore of the magnet.

The RF transmit coil generates appropriately shaped pulses of current at the Larmor frequency, producing an alternating B_1 field. The weak MR signal resulting from transverse magnetization is detected by receiver coils. There are two types of receiver coil: volume and surface. Volume coils completely encompass the region of interest and are often used as combined RF transmit/receive coils. Surface coils are specifically designed for different body sites and are placed close to the surface of the patient. These coils are generally receive-only due to their inhomogeneous reception field, and their sensitivity is dependent on the shape and arrangement of the individual coil elements.

A Faraday cage encloses MRI scanner room to minimize RF contamination. Outside the Faraday cage, computer systems are housed in a separate control room, where pulse sequence prescription, signal processing, and image reconstruction are performed.

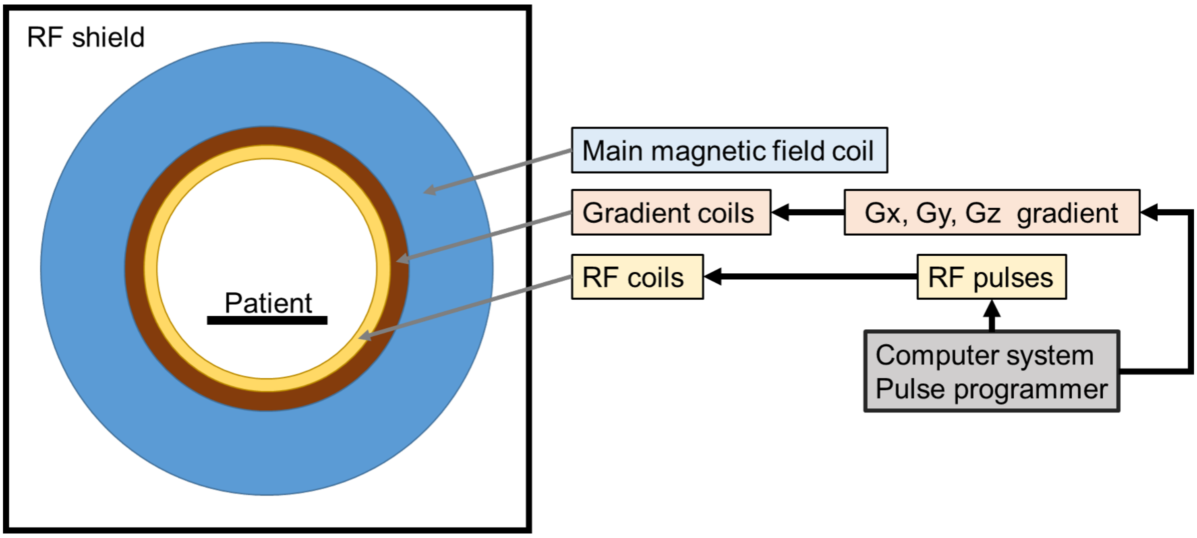

Figure 2.5: Schematic representation of the fundamental components of an MR scanner.

2.8.2 Linear accelerator components

Fig. 2.6 illustrates the main technical components of a representative medical linear accelerator. The electron gun generates electrons through thermionic emission, where a heated cathode provides sufficient energy for electrons to overcome the material's work function. The electrons are then directed into the circular waveguide to be accelerated by the RF pulses to a speed $v \approx c$, with c being the speed of light [132]. The RF waves are produced by a magnetron with an electromagnetic field. The kinetic energy of the electrons E_{kin}, is related to the voltage U applied in the magnetron, as follows:

$$E_{kin} = e \cdot U \tag{2.31}$$

where e is the elementary charge of an electron. After exiting the waveguide, the electrons enter the flight tube, where the beam is bent and focused by a set of magnets to hit the target. The high-energy electron beam strikes a small tungsten target and emits photons via Bremsstrahlung and characteristic radiation. The resulting X-ray beam has an energy range of 4 to 25 megavolts (MV).

The ionization chamber is integrated to monitor the amount of radiation that passes through. Following beam shaping and size-limiting by the primary collimator, the multi-leaf collimator, composed of tungsten leaves, further shapes the radiation beam to conform to the tumor contour and enables the decomposition of the treatment field into smaller subfields. The gantry rotates around the patient, with the isocenter defined as the intersection of the gantry's rotation axis and the central axis of the radiation beam.

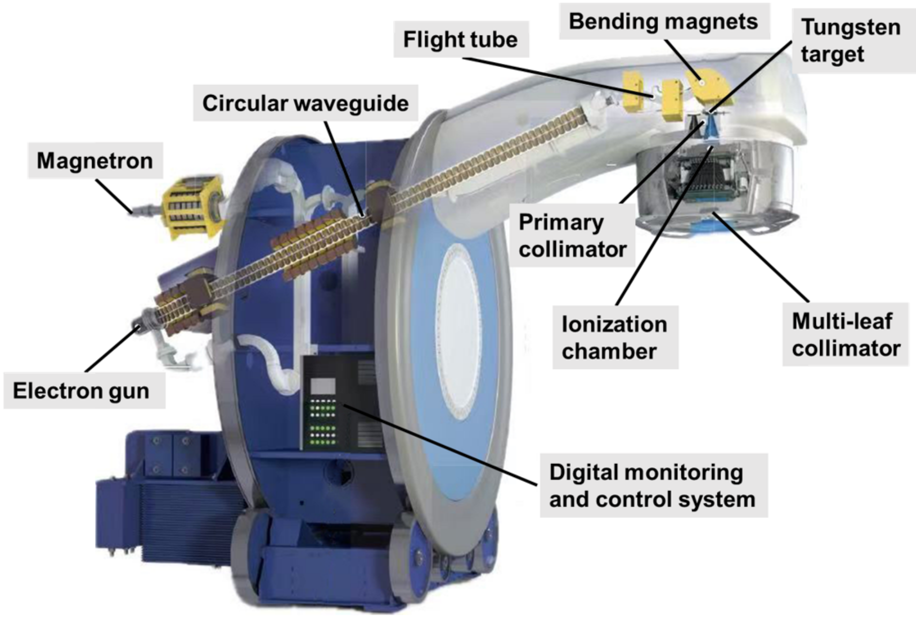

Figure 2.6: Main technical components of a representative medical Linac from Elekta [133]. Figure adapted from [134] with permission.

2.8.3 Integration of MRI and Linacs: the MR-Linac

The mutual interference between a strong magnetic field and the field-sensitive components of the linear accelerator presents a major challenge in integrating an MRI system with a Linac into a single treatment device.

The effects of a magnetic field on Linac operation can be summarized as follows [70]: a standard clinical Linac has a magnetic field tolerance of 1 G (0.0001 T), whereas the MRI system operates at field strengths of up to 1.5 T, posing a significant challenge to the functionality of field-sensitive components. The primary concern is the performance of the magnetic encoders in the MLC, which control the positioning of motor-driven leaves. Additionally, the magnetic field can induce deviations of electrons within the waveguide, leading to beam current loss. Moreover, the Lorentz force on the secondary electrons—generated through photon interactions with matter—can alter their paths. Specifically, when the static magnetic field is parallel to the radiation beam, this results in an electron focusing effect (EFE), whereas a perpendicular orientation leads to an electron return effect (ERE). These phenomena can potentially impact dose distribution in specific tissues.

Conversely, Linac components also impact the MRI operation [70]: The RF noise and heterogeneity of the main magnetic field, caused by the proximity of MLC, can degrade image quality and cause geometric distortions. Additionally, the RF receiver coil in the path of the beam can attenuate the intended dose while increasing skin dose via secondary electrons. It can also induce electronic disequilibrium in the conductors or electronics, resulting in imaging artifacts.

Addressing these challenges requires dedicated efforts, including active shielding, reduction in field strength, implementation of field-compatible components, and incorporation of Lorentz forces effects in dose calculations. Existing MR-Linac systems demonstrate that different design strategies can be adopted to achieve an integrated MRgRT delivery system [135]. Fig. 2.7 illustrates a solution from Viewray MRIdian MR-Linac, which employs a low-field MRI scanner at 0.35 T with a split superconducting double-donut magnet. This system is installed at LMU University Hospital, where it is used for both clinical treatments and research. The circular radiation gantry is positioned within the gap between the magnet halves, allowing the treatment beam to be emitted perpendicular to B_0. Six shielding compartments are mounted on the gantry to protect the internal Linac components and MLC from the magnetic field, while also providing RF shielding for MR imaging during the Linac operation. Both the MRI and Linac share the same isocenter.

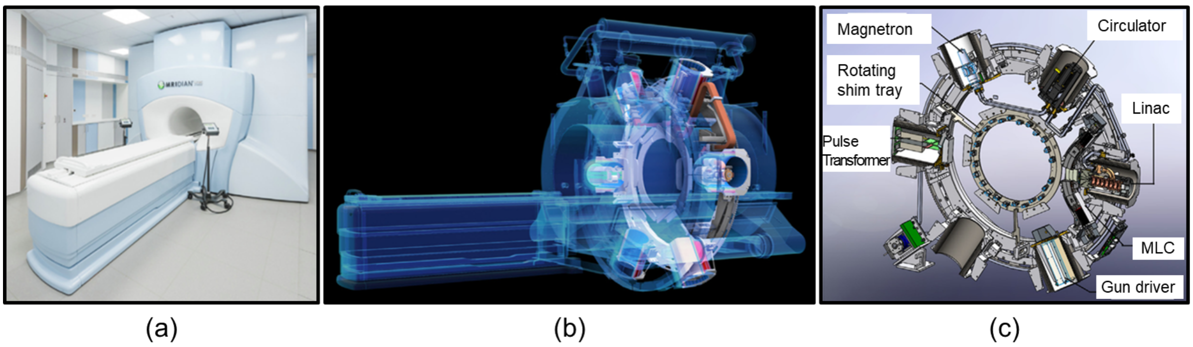

Figure 2.7: (a) Photograph of the ViewRay MRIdian MR-Linac system. (b) Schematic drawing of the system depicting the main hardware components: superconducting double-donut magnet, circular radiation gantry and patient couch. (c) Schematic drawing of the radiation gantry with linac components and MLC. Images courtesy of ViewRay Inc. Figure adapted from [136].

Beyond addressing the mutual interference between the MRI magnetic field and field-sensitive Linac components, the most critical challenge in this complex MR-Linac system is geometric errors arising from latency, with MR imaging latency identified as the dominant contributor. This work focuses on mitigating these MR imaging errors during real-time motion monitoring in MRgRT through deep learning-based approaches. In the following chapters, the motion-dependent k-space sampling process is first simulated, followed by the introduction of the dataset creation methods for machine learning

(Chapter 3). Chapters 4 and 5 then present solutions tailored for Cartesian and radial readout trajectories, respectively.

Chapter 3

SIMULATION OF MOTION-RELATED IMAGING ERRORS AND DIGITAL PHANTOM-BASED DATASET CREATION

3.1 Simulation of motion-related imaging errors

As outlined in Chapter 2, the raw MR signal is sampled in the frequency domain, with k-space being progressively filled during the acquisition process. Since the acquisition of a single cine-MR frame occurs over a duration comparable to the physiological motion timescale, the corresponding k-space becomes populated with signals originating from multiple target positions. When transformed into the spatial domain, these temporally mixed signals give rise to motion-related imaging errors, which are a critical concern in the context of real-time motion monitoring during MRgRT.

To effectively mitigate such errors, an in-depth understanding of how the moving target is captured, or how the intra-frame motion is encoded, during MRI acquisition—specifically in the raw k-space signal and its subsequent translation into image space—is essential. Therefore, the motion-dependent k-space sampling process is simulated in this section to facilitate a fundamental understanding of this dynamic MR imaging behavior and to elucidate the origin and characteristics of motion-related imaging errors. A dedicated simulation platform has been developed and validated, further serving as the foundation for generating datasets suitable for training deep learning-based intra-frame compensation models, as detailed in the subsequent sections.

3.1.1 Development of the simulation platform

Fig. 3.1 schematically summarizes the motion-dependent sampling in the simulation procedure. In the Amplitude/Time curve, the black dots indicate the start/end time of acquiring a specific cine-MR frame and the corresponding positions. The dotted line connecting these black dots represents the intra-frame motion trajectories of the target. Assume that the frame acquisition time is divided into *ns* steps (or shots), with each step (shot) corresponding to a segment of k-space that can be approximated as being acquired simultaneously, such as a spoke in the radial trajectory or a phase-encoding line in the Cartesian trajectory (Section 2.5). The simulation procedure consists of three main modules: determination of the temporal images, definition of the k-space readout trajectory, and reconstruction of the motion-corrupted image.

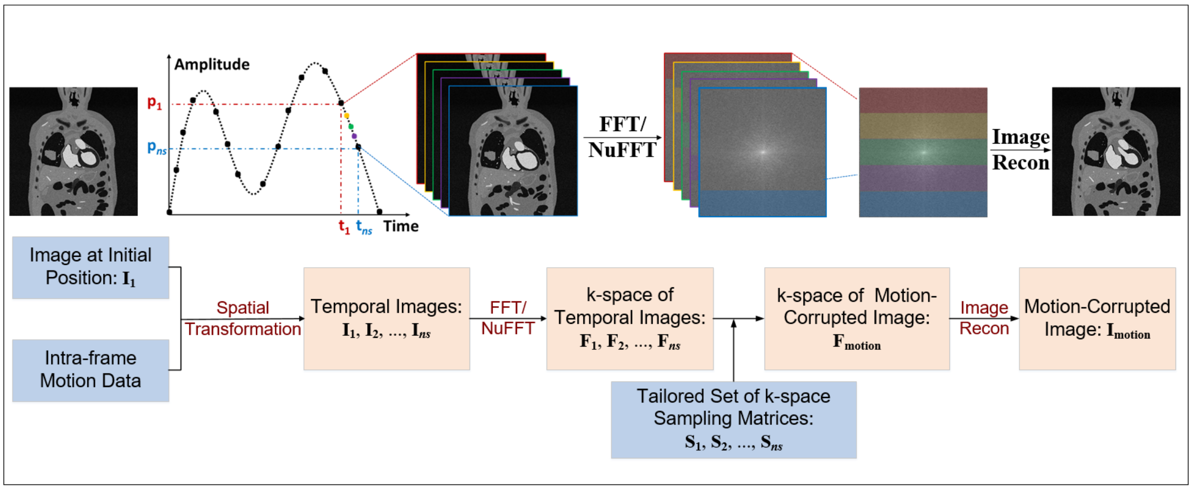

Figure 3.1: A schematic diagram of the motion-dependent k-space acquisition simulation procedure. This diagram takes Cartesian sampling as an example, with 5 temporal segments (*ns*=5) for easier visualization. In actual applications, a significantly larger number of shots is employed. This figure was originally published in [137].

Determination of the temporal images

Assume that the acquisition of a frame begins at time t_1, with the target initially positioned at p_1. During the acquisition period, the target transitions through various intermediate positions, ultimately reaching its final position p_{ns} by the time the acquisition concludes at t_{ns}. The images corresponding to these time steps are referred to as temporal images ($\mathbf{I_1}, \mathbf{I_2}, \ldots, \mathbf{I}_{ns}$), capturing ground-truth discrete anatomical positions.

Intra-frame motion data can be represented in various forms: including translation or rotation parameters for rigid motion, control point movements for free-form deformation (FFD), or, more commonly, displacement vector fields (DVF) for image warping, where the displacement of each pixel or voxel in the image is specified. Motion at each time step is characterized individually, resulting in a series of parameters for the entire intra-frame motion sequence. Starting from the image at the initial position ($\mathbf{I_1}$) or the final position ($\mathbf{I}_n$) and utilizing the intra-frame motion data, the temporal images of the target throughout the acquisition period can be established or derived by image spatial transformations.

Definition of the k-space readout trajectory

In alignment with the k-space readout patterns commonly used in clinical MR-Linac systems, this study considers both Cartesian and radial acquisitions.

For Cartesian acquisitions, the phase or frequency encoding direction is first specified according to the requirements. The user can then customize the k-space profile ordering as needed along the phase-encoding axis. Fig. 3.2 presents several examples of Cartesian phase-encoding ordering schemes with partial Fourier [138] acceleration: *linear*, where encoding proceeds from bottom to top; *reverse-linear*, where encoding proceeds from top to bottom; and *high-low*, where, based on the frequency order, higher frequencies are acquired first, followed by lower frequencies.

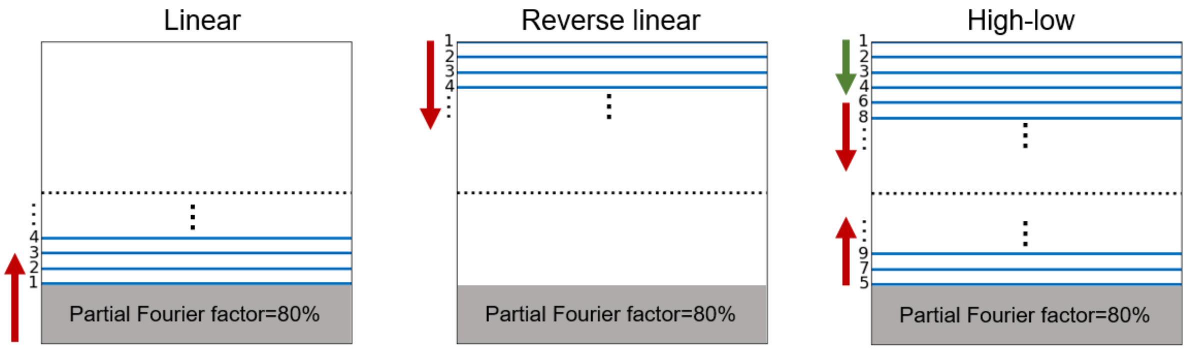

Figure 3.2: Schematic diagram of exemplary Cartesian phase-encoding ordering schemes with a partial Fourier factor of 80%. *linear*: encoding proceeds from bottom to top; *reverse-linear*: encoding proceeds from top to bottom; and *high-low*: based on the frequency order in the phase encoding direction, higher frequencies are read first, followed by lower frequencies.

As illustrated in Fig. 3.3, in radial trajectories, all spokes pass through the origin, with the first spoke forming an angle of γ with the horizontal axis. Subsequent spokes are placed using a successive azimuthal increment of ψ. Following real-world applications, three types of radial trajectories are simulated and investigated in this study based on the value of ψ: *linear*, *golden angle*, and *tiny golden angle*.

Radial sampling with a *linear* profile ordering takes $\psi = \psi_{\text{linear}} = \pi/ns$, covering k-space uniformly only after acquiring the complete set of ns spokes. To achieve a nearly uniform profile distribution in k-space for an arbitrary number of radial spokes, *golden angle* trajectories were proposed [139], employing

$$\psi = \psi_{\text{gold}} = \pi/\tau \tag{3.1}$$

where $\tau = \left(1 + \sqrt{5}\right)/2$ represents the golden ratio. However, the large azimuthal profile increment required by this method can cause strong eddy current artifacts due to rapid gradient switching [140]. To address this issue, a surrogate *tiny golden angle* profile ordering, defined by

$$\psi = \psi_N = \pi/(\tau + N - 1) \quad \text{where} \quad N = 3, 4, \ldots \tag{3.2}$$

was introduced [141] and has found widespread use in real-time MR imaging [142].

In the simulation procedure, the readout trajectories for radial sampling are generated by defining the coordinates of each uniformly spaced sampling point on each spoke using γ and ψ:

$$\begin{aligned} u(s,r) &= \left(\frac{2\pi(r-1)}{N_{\text{point}}} - \pi\right) \cdot \cos\left(\gamma + (s-1)\psi\right), \\ v(s,r) &= \left(\frac{2\pi(r-1)}{N_{\text{point}}} - \pi\right) \cdot \sin\left(\gamma + (s-1)\psi\right), \\ \text{for} \quad s &= 1, 2, \ldots, N_{\text{spoke}} \quad \text{and} \quad r = 1, 2, \ldots, N_{\text{point}} \end{aligned} \tag{3.3}$$

where $u(s,r)$ and $v(s,r)$ represent the horizontal and vertical coordinates of the r-th sampling point on the s-th spoke, respectively; N_{spoke} and N_{point} denote the total number of spokes and the number of points along each spoke, respectively.

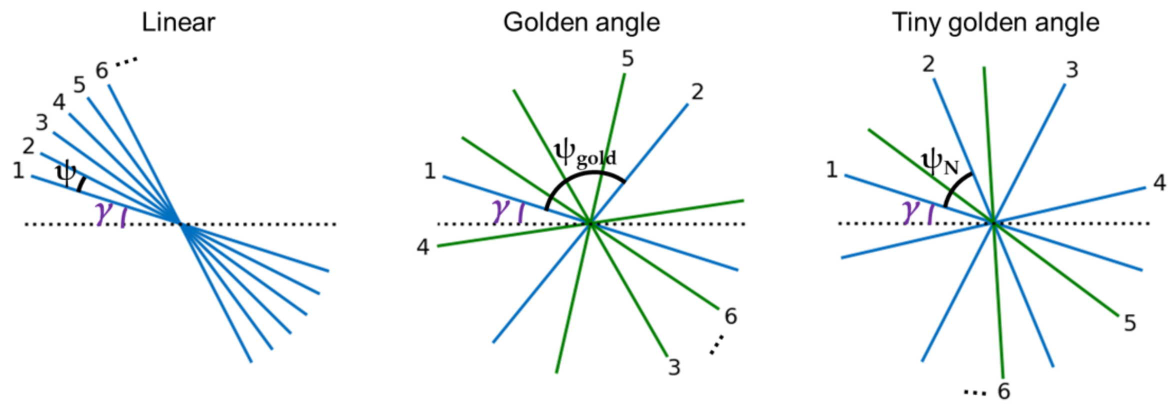

Figure 3.3: Schematics of radial sampling trajectories. From left to right: *linear*, *golden angle*, and *tiny golden angle*. γ denotes the initial spoke angle; ψ, ψ_{gold}, and ψ_N indicate the azimuthal increments of the radial profiles for the respective trajectories.

Reconstruction of the motion-corrupted image

Depending on the k-space sampling trajectories, either an FFT for Cartesian readouts or a NuFFT for radial readouts (see Section 2.7.1) is performed to obtain the complex-valued k-space data ($\mathbf{F_1}$, $\mathbf{F_2}$, ..., $\mathbf{F}_{ns}$) of each temporal image. Their corresponding components are then sequentially incorporated into the k-space arrays of the simulated motion-corrupted image over time. Specifically, the first shot extracts the $\frac{1}{ns}$-th components of the k-space from the initial-position image. Subsequent shots extract the $\frac{i}{ns}$-th components of the k-space corresponding to the i-th temporal image, up to the final time step, where the last components are extracted from $\mathbf{F}_{\boldsymbol{ns}}$. This process can be achieved by designing a tailored set of sampling matrices ($\mathbf{S_1}$, $\mathbf{S_2}$, ..., $\mathbf{S}_{\boldsymbol{ns}}$) based on the k-space readout trajectories with respect to the shot number, where MRI acceleration techniques like partial Fourier or parallel imaging methods [127, 138] should be considered. These sampling matrices have only two values, 0 and 1, with 1 indicating that the corresponding part of the matrix will be sampled. Consequently, the complex-valued k-space of the motion-corrupted image $\mathbf{F_{motion}}$ is formulated as:

$$\mathbf{F_{motion}} = \sum_{j=1}^{ns} \mathbf{F}_j \circ \mathbf{S}_j = \sum_{j=1}^{ns} \mathcal{F}_2\left(\mathbf{I}_j\right) \circ \mathbf{S}_j \tag{3.4}$$

where *j* denotes the shot number; $\circ$ is the Hadamard product (element-wise product) operator; and $\mathcal{F}_2$ is the 2D Fourier transform operator (FFT for Cartesian and NUFFT for radial, see Section 2.7). In accordance with the real-world conditions, the finally acquired k-space $\mathbf{F_{motion}}$ consists of signals from the target at different positions.

Finally, the motion-corrupted image $\mathbf{I_{motion}}$ is reconstructed as a complex-valued image using relevant image reconstruction techniques, such as inverse FFT/NuFFT, GRAPPA, and others, as detailed in Section 2.7:

$$\mathbf{I_{motion}} = \mathcal{F}_2^{-1}\left(\mathbf{F_{motion}}\right) = \mathcal{F}_2^{-1}\left(\sum_{j=1}^{ns} \mathcal{F}_2\left(\mathbf{I}_j\right) \circ \mathbf{S}_j\right) \tag{3.5}$$

where $\mathcal{F}_2^{-1}$ denotes the 2D inverse Fourier transform operator.

Differences between $\mathbf{I_{motion}}$ and the ground-truth final-position image $\mathbf{I}_{\boldsymbol{ns}}$ reflect the intra-frame motion deterioration effects, that is, the motion-related imaging errors.

3.1.2 Validation of the simulation platform

3.1.2.1 Validation based on theoretical analysis

The motion-dependent k-space sampling simulation procedure was first validated through a theoretical analysis based on the translational and rotational properties of the Fourier transform, with the corresponding proof provided in the Appendix A. The translational property indicates that shifting an image $f(x, y)$ by Δx in the x-direction and by Δy in the y-direction induces a linear phase shift in the corresponding Fourier domain, while the magnitude of the Fourier transform remains unchanged. Let $\hat{F}(u, v)$ and $F(u, v)$ denote the Fourier transforms of the translated and original images, respectively. This relationship can be expressed mathematically as:

$$\hat{F}(u,v) = F(u,v)\exp\left[-j2\pi(u\Delta x + v\Delta y)\right] \tag{3.6}$$

Furthermore, the rotational property asserts that a rotation in the spatial domain corresponds to a rotation by the same angle in the frequency domain.

Leveraging these properties of the Fourier transform, a digital cross phantom of size 140×140 pixels was constructed for validation purposes and mathematically defined as:

$$\begin{gathered} f(i,j) = \chi_{[35,105)}(i)\cdot\delta_{j,70} + \delta_{i,70}\cdot\chi_{[35,105)}(j) - \delta_{i,70}\cdot\delta_{j,70} \quad ; \\ i,j \in \{0,1,2,\ldots,140\} \end{gathered} \tag{3.7}$$

where $\chi_{[m,n)}(x)$ is the indicator function for the interval $[m,n)$, and $\delta_{x,a}$ is the Kronecker delta function, defined as:

$$\chi_{[m,n)}(x) = \begin{cases} 1 & \text{if} \quad m \le x < n, \\ 0 & \text{otherwise.} \end{cases} \quad ; \quad \delta_{x,a} = \begin{cases} 1 & \text{if} \quad x = a, \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

The image and k-space of the cross phantom are presented in the "Initial-position" column of Fig. 3.4. Notably, the magnitude of k-space is non-zero throughout, displaying a high-contrast horizontal and vertical line that are orthogonal to the origin.

In Fig. 3.4, the cross phantom was rotated 45° counterclockwise from its initial position during frame acquisition. The simulated motion-dependent sampling process was divided into $ns = 10$ shots, with each shot rotated by 5°. A comparison between the final-position image and the initial-position image reveals that the k-space was also rotated 45° counterclockwise. Additionally, it is observed that the motion-corrupted k-space sequentially captured segments of the Fourier data from the temporal images, where each image was sequentially rotated by 5° counterclockwise relative to the shot number, in both the spatial and frequency domains. This finding is consistent with theoretical expectations.

The simulation procedure was further validated by examining the imaging behavior of the cross phantom's translation, where the relationship between $\hat{F}(u,v)$ and $F(u,v)$, as described in Eq. 3.6, was verified. Since the phase of the complex number lies within the interval $(-\pi,\pi]$ and to avoid phase wrapping, rather than directly comparing the phase or magnitude of the two terms, a transfer matrix $M(u,v)$ was defined for the cross phantom as:

$$M(u,v) = \frac{\hat{F}(u,v)}{F(u,v)} \tag{3.9}$$

where $F(u,v)$ was substituted with the complex-valued k-space data of the initial-position phantom image, which was non-zero throughout; and $\hat{F}(u,v)$ was replaced by the corresponding k-space data of either the final-position image or the simulated motion-corrupted image. The magnitude and phase of $M(u, v)$ were then computed as the modulation transfer matrix and phase transfer matrix, respectively.

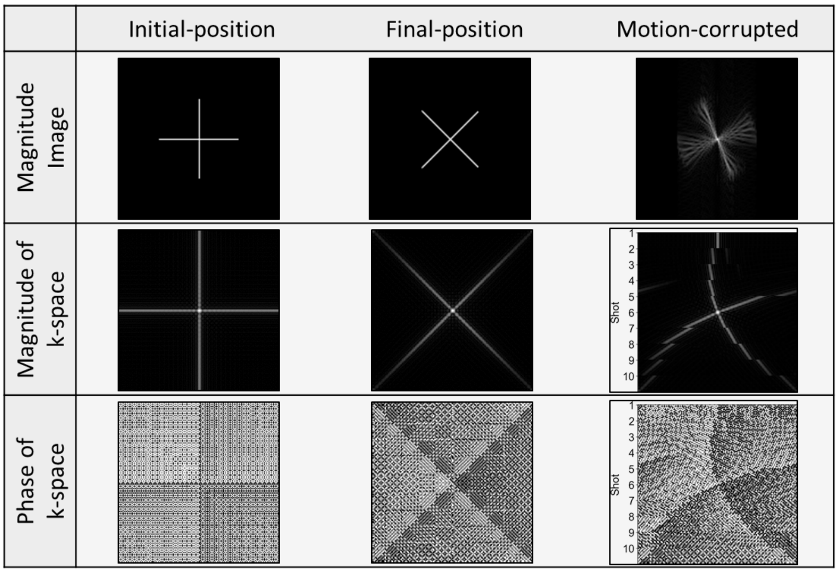

Figure 3.4: Rotation experiment of the cross phantom. The phantom has been rotated 45° counterclockwise from its initial position during frame acquisition, with $ns = 10$. The figure displays images at the initial and final positions, as well as the simulated motion-corrupted image, in both the spatial and frequency domains; the phase encoding direction is vertical (up-down).

Fig. 3.5 presents the results of the translation experiments. To minimize potential interpolation errors, each shot was configured to move the phantom by exactly one pixel. In the first two experiments, the phantom was shifted 4 pixels downward and 4 pixels to the right from its initial position, respectively, with the simulated motion-dependent sampling process divided into $ns = 5$ shots. In the third experiment, the phantom was moved 9 pixels both downward and to the right simultaneously, with $ns = 10$.

The results indicate that, aside from computational precision limits, the values of the modulation transfer matrix for both the final-position image and the motion-corrupted image were equal to 1 across all experiments. Moreover, the phase transfer matrix derived from the final-position image exhibited a periodic pattern in the direction of movement, with the number of phase cycles matching the displacement magnitude. Specifically, Experiment 1 displayed 4 cycles of $(-\pi, \pi]$ in the vertical direction, and Experiment 2 showed 4 cycles in the horizontal direction; whereas in Experiment 3, the pattern revealed 9 cycles in both the vertical and horizontal directions simultaneously. Nevertheless, the frequency-domain information of the temporal images had been spatially and temporally encoded in the motion-corrupted k-space. The accumulation of displacement over the acquisition time steps, reflected by the progressively increasing number of cycles along the shift direction within the phase transfer matrix, is clearly represented in the motion-corrupted image. This observation aligns with what was theoretically anticipated based on Eq. A.4.

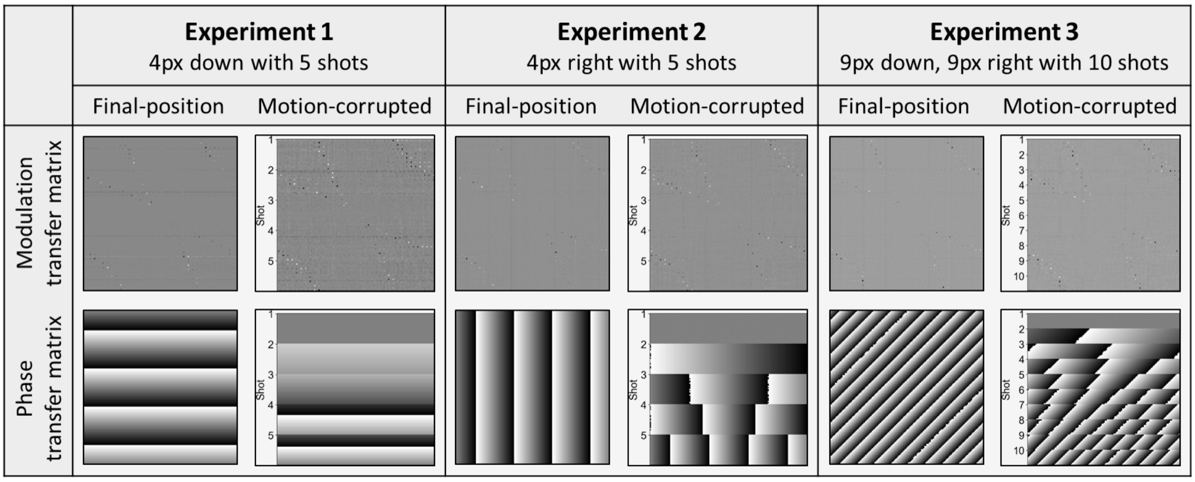

Figure 3.5: Translation experiments of the cross phantom. From left to right: Experiment 1, the phantom is shifted 4 pixels downward with $ns = 5$; Experiment 2, the phantom is shifted 4 pixels to the right with $ns = 5$; Experiment 3, the phantom is shifted 9 pixels downward and to the right simultaneously, with $ns = 10$. The modulation and phase transfer matrices derived from both the final-position image and the motion-corrupted image are displayed; the phase encoding direction is vertical (up-down).

3.1.2.2 Validation against literature: Imaging latency experiments

In the work by Borman and colleagues [88], the target positioning errors were characterized as imaging latency, further identified as the largest contributor to the total system latency in MRgRT [32, 33]. They measured MR imaging latency through simulations and experiments using an MR-compatible motion platform (ModusQA,CA) in a 1.5T MR-linac and a 3T MR scanner. By synchronizing machine-acquired images with the the physical motion platform, and fitting the sinusoidal model, $x(t) = A_0 \sin(2\pi f[t + t_0])$, to both the reference and MR-derived position traces, they estimated the latency for Cartesian and radial acquisitions with various k-space profile orderings. In this section, the proposed simulation procedure is validated against the findings of these imaging latency experiments reported in the literature [88], ensuring that it is closely aligned with real-world conditions.

Following Borman and colleagues' work, where the simulated imaging latency was determined by tracking the movement of an analytical square during k-space filling, a digital square phantom was constructed, as shown in Fig. 3.6 (a). Figs. 3.6 (b) and (c) illustrate examples of motion-dependent Cartesian acquisition with a vertical (up-down) phase encoding direction, where the phantom moved downward and to the right, respectively. Compared to motion in a perpendicular direction, the motion-corrupted image appeared slightly more blurred when the motion was parallel to the phase encoding direction. However, target positioning errors were more pronounced, emerging as the dominant factor in imaging errors. The center of mass (COM) of the target was computed from the motion-corrupted images as the measured position, while the reference position was determined from the actual target COM at the end of the frame acquisition, corresponding to the final shot. By analyzing the target motion speed and computing the distance between the measured and reference positions, the imaging latency was estimated.

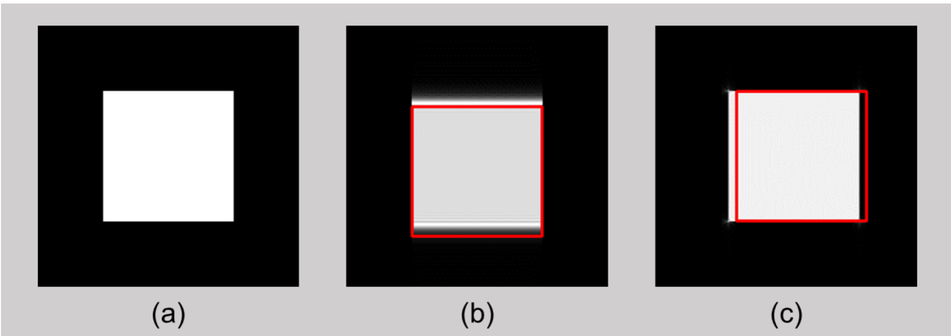

Figure 3.6: Square phantom for imaging latency experiments. (a) Phantom at the initial position. Examples of simulated motion-corrupted images in the Cartesian experiments are shown for the phantom moving downward (b) and to the right (c). The red contours indicate the actual target positions at the end of the frame acquisition (corresponding to the last shot); the phase encoding direction is vertical (up-down).

The Cartesian experiments were carried out with a range of partial Fourier factors, defined as the ratio of sampled data to the full matrix, across the three phase-encoding ordering schemes depicted in Fig. 3.2. The k-space phase encoding direction was orthogonal to the primary direction of target motion. Fig. 3.7 compares the results obtained from the literature [88] with those generated by the motion-dependent sampling simulator developed in this study, revealing close agreement. For a given partial Fourier factor and k-space profile ordering scheme, the ratio of imaging latency to frame acquisition time remained relatively consistent across all experiments. Prior studies have shown that, for Cartesian readout trajectories, the object position is primarily determined by the moment at which the central k-space profile is acquired [143]. The simulation results reproduced this behavior, demonstrating that imaging latency can be approximated as the time difference between the shot of acquiring the central k-space profile and the last shot. Specifically, the *linear* phase-encoding ordering yields similar imaging latency regardless of the partial Fourier factor; however, with the *reverse linear*, the ratio of imaging latency to frame acquisition time remains unchanged as the partial Fourier factor increases, thus showing a latency proportional to the acquisition time; furthermore, the *high-low* scheme leads to the minimum imaging latency since the central k-space profile is always the last to be acquired.

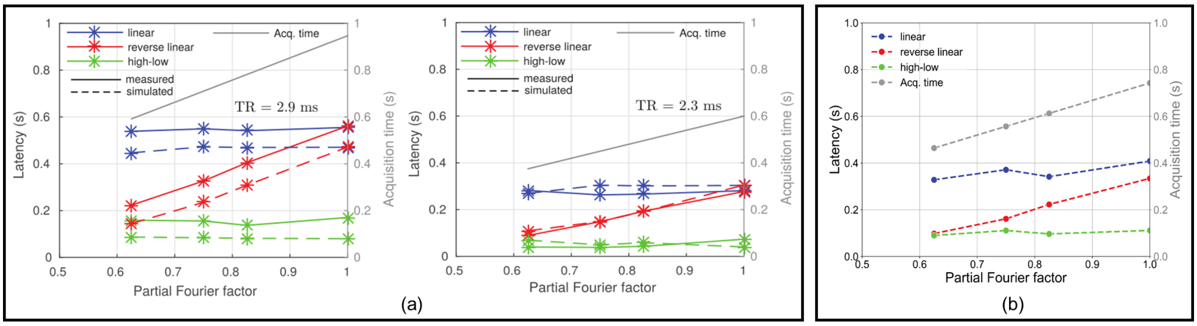

Figure 3.7: Results of Cartesian imaging latency experiments. (a) from Borman et al. [88], featuring the 1.5T MR-Linac (left) and 3T MR scanner (right); (b) from the simulation procedure developed in this study.

The radial experiments were conducted using *linear* and *golden angle* profile orderings (illustrated in Fig. 3.3), with the corresponding results presented in Fig. 3.8 and Fig. 3.9. To vary the sampling fractions, the number of radial spokes was adjusted proportionally. The oversampled central k-space of radial sampling, along with the nearly uniform k-space coverage provided by the *golden angle* profile ordering, allows for the application of various k-space data weighting schemes [144, 145]. To reduce the radial imaging latency, Borman et al. implemented a spatial-temporal (k-t) filter that attenuated the low-frequency components of previously acquired spokes while preserving the high-frequency components [88]. The k-t filter was defined as a function of the spoke index s and the readout point k, and was mathematically expressed as:

$$f(k, s) = [1 - S(s)]k^2 + S(s), \text{ where } S(s) = \frac{1}{1 + e^{-\alpha(s - n_s/2)}} \tag{3.10}$$

Here, n_s was the total number of spokes, and α was a parameter determined via hyperparameter search. The filter employed a sigmoid function to distinguish between earlier and later sampling of the spokes, while a quadratic function applied high-pass weighting to the readout points along each spoke. For comparison, this study also conducted experiments using this same filter.

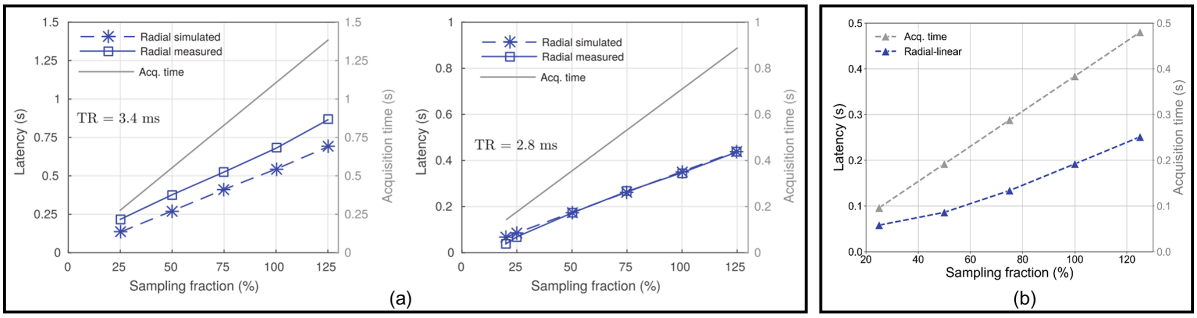

Figure 3.8: Results of radial imaging latency experiments with *linear* profile orderings. (a) from Borman et al. [88], featuring the 1.5T MR-Linac (left) and 3T MR scanner (right); (b) from the simulation procedure developed in this study.

Consistent with the Cartesian experiments, the radial imaging latency estimated using the proposed procedure closely aligns with the findings reported in the literature. The latency was approximately 50% of the total frame acquisition time, indicating that each radial spoke plays an equal role in determining the target position within the image. Fig. 3.9 demonstrates that retrospectively weighting the radial k-space data with a k-t filter for *golden angle* sequences, as defined by Eq. 3.10, can effectively halve the imaging latency.

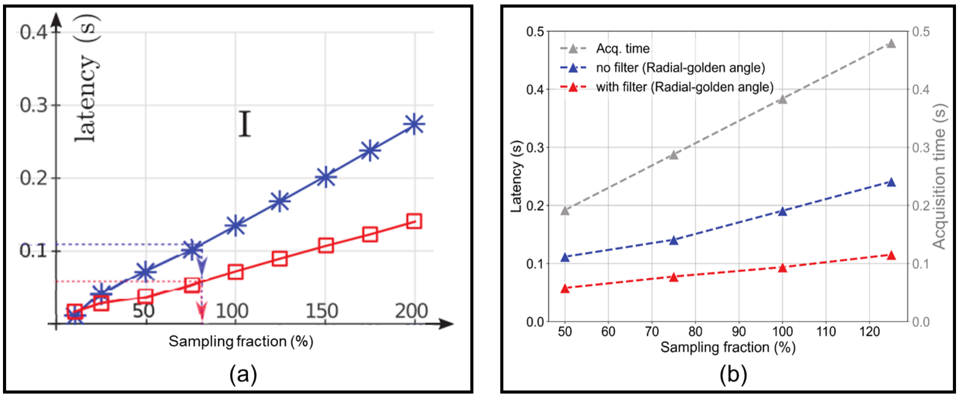

Figure 3.9: Results of radial imaging latency experiments with *golden angle* profile orderings, with (red) and without (blue) k-t filter. (a) Adapted from Borman et al. [88], measured on a 3T MR system; (b) Obtained from the simulation procedure developed in this study.

The results of both Cartesian and radial experiments validated the accuracy of the motion-related imaging error simulation procedure developed in this study.

3.2 Formulation of the inverse problem and deep learning solution for intra-frame motion compensation

The observation of motion-related imaging errors underscores the practical significance of compensating for intra-frame motion, particularly in cases involving rapid anatomical variations.

The motion-dependent k-space acquisition process, together with the simulated motion-corrupted images and k-space representation in Fig. 3.4 and Fig. 3.5, demonstrates that part of the frequency domain information from each temporal image is spatially and temporally encoded within the k-space of the obtained motion-corrupted image. The encoding is uniquely dictated by the predefined k-space readout trajectory. As a result, an inverse problem can be formulated to recover the implicit real-time final-position image (i.e., the last-shot temporal image) from the motion-corrupted image or its k-space, thereby compensating for the intra-frame motion:

$$\begin{aligned} \mathbf{I}_{ns} &= \mathscr{T}_I\left(\mathbf{I}_{\text{motion}}\right) \\ \mathbf{F}_{ns} &= \mathscr{T}_F\left(\mathbf{F}_{\text{motion}}\right) \end{aligned} \tag{3.11}$$

where $\mathscr{T}_I$ and $\mathscr{T}_F$ denote the transformations that aim to derive the final-position image and k-space, respectively, from their motion-corrupted counterparts. Owing to the information loss and non-uniqueness induced by intra-frame motion-dependent k-space acquisition, the problem is inherently ill-posed. To obtain a stable and reliable solution, appropriate regularization terms or learned priors should be incorporated.

A neural network, by learning a mapping function between the input and output spaces, provides a compelling data-driven solution to address the inverse problem. In particular, supervised learning is widely adopted for such tasks, where the network is trained to minimize the discrepancy between predicted and reference outputs. Accordingly, the optimization problem associated with Eq. 3.11 is given by:

$$\begin{aligned} \min_{\theta_I} \ \sum_i \mathcal{L}\left(\mathcal{N}_I(\mathbf{I}^{(i)}_{\text{motion}};\theta_I), \mathbf{I}^{(i)}_{ns}\right) \\ \min_{\theta_F} \ \sum_i \mathcal{L}\left(\mathcal{N}_F(\mathbf{F}^{(i)}_{\text{motion}};\theta_F), \mathbf{F}^{(i)}_{ns}\right) \end{aligned} \tag{3.12}$$

where $\mathcal{L}$ denotes the loss function (e.g., ℓ_1 or ℓ_2 norm). $\mathcal{N}_I$ and $\mathcal{N}_F$ denote the neural networks that approximate the transformations $\mathscr{T}_I$ and $\mathscr{T}_F$ in Eq. 3.11, respectively. θ_I and θ_F represent the trainable weights of the networks, and i indexes the training samples.

Through deep learning, the network can be trained to extract relevant information from the later-acquired portions of motion-corrupted data and leverage this information to correct earlier acquired components, thereby aligning the reconstructed image with the target's final-position reference. Given the unique characteristics of Cartesian and radial k-space readout trajectories, the network architecture must be specifically tailored to accommodate each sampling pattern. Moreover, constructing a suitable training dataset is crucial to ensure the network's effectiveness.

3.3 Motivation for creating datasets using simulated phantoms

To enable supervised data-driven learning, the creation of labeled datasets is essential. In this study, this involves pairing each motion-corrupted image with its corresponding ground-truth final-position image for the moving target.

However, in clinical practice, cine-MR frames are often already contaminated by the intra-frame motion of the target, leading to errors in target positioning and shape representation. Determining the ground truth for motion-related imaging error reduction in the clinic is more challenging compared to other AI application scenarios in MRgRT, such as image segmentation, where training pairs consist of clinically acquired images as inputs and ground-truth contours, generated and approved by radiation oncologists, as outputs [114]. This increased difficulty arises because imaging errors are often harder for domain experts to detect than segmentation errors [90]. A similar setup to synchronize the machine-acquired images with the physical motion platform may be required, as implemented in Borman et al.'s work [88]. Nonetheless, typical MR motion phantoms [146] are often overly simplistic in geometry, and are restricted to the rigid motion of small targets, which is inadequate for building comprehensive training datasets. More suitable alternatives include anthropomorphic phantoms or relatively complex phantoms, such as the porcine lung phantom [147].

While the physical MRI-compatible anthropomorphic moving phantom incurs significant costs in both time and financial investment, and the complexity of clinical experiments demands considerable and dedicated efforts, digital phantoms have emerged as a practical solution to address the lack of in vivo ground truth [148, 149]. This is supported by the following considerations:

Firstly, the signal acquisition simulator can be designed to closely replicate real machine conditions. As demonstrated in the previous section (Section 3.1.2.2), the imaging latency results obtained from the simulation procedure developed in this study show negligible differences compared to those reported by Borman et al. [88] in clinical experiments conducted on the 1.5T MR-Linac and 3T MR scanner.

Secondly, compared to clinical experimental data, simulated data offer precise final-position images and target segmentation for ground truth and evaluation, remaining unaffected by other sources of imaging uncertainty.

Thirdly, sufficient spatial resolution is a prerequisite for studying intra-frame motion. Digital phantoms overcome the spatial resolution limitations typically encountered in clinical cine-MR images, which may be insufficient for investigating positioning accuracy. They also enable complex-valued image reconstruction with various dedicated k-space readout trajectories and noise models, facilitating exploratory research.

Finally, as indicated in the literature, respiratory motion exhibits patient-specific characteristics, making it unpredictable, irregular, and subject to temporal variation. Variations influenced by tumor location and pathology result in unique patterns of displacement, direction, and motion phase. This study focuses particularly on rapid and uncommon anatomical changes. In real-world scenarios, most cases involve small or moderate motion; nevertheless, although rapidly moving targets are less common, they require compensation more urgently. Simulated data facilitates the creation of deep breathing motion scenarios and enables the customization of arbitrary motion patterns, both of which are highly relevant and central to the study of intra-frame motion compensation. Additionally, extreme motion scenarios, which are uncommon or unlikely in real-life conditions, can be incorporated to introduce a significant deviation between the network's input and output, forcing the model to focus on dynamic mechanisms and avoid potential extrapolation errors related to motion amplitude.

Therefore, to conduct a proof-of-concept study on the deep learning-based intra-frame motion compensation technique and to demonstrate its feasibility and real potential in reducing cine-MR imaging errors, simulated data will initially be utilized for dataset creation, facilitating some principle results and evidence.

3.4 Digital phantom-based dataset creation

High-quality datasets play a critical role in the successful application of deep learning-based techniques. This section outlines the process of creating labeled datasets specifically for deep learning-based intra-frame motion compensation.

The primary consideration when establishing the database is the development and integration of various types of motion data. Fig. 3.10 illustrates the two-scale discretization of the motion trajectory throughout this work, capturing anatomical variations at both coarse and fine levels of granularity. This coarse-to-fine strategy underpins the two main steps in the dataset creation process: (i) the generation of 4D MRI digital anthropomorphic phantoms to represent key anatomical positions during breathing, and (ii) the synthesis of intra-frame motion data for determining temporal images, as described in Section 3.1.

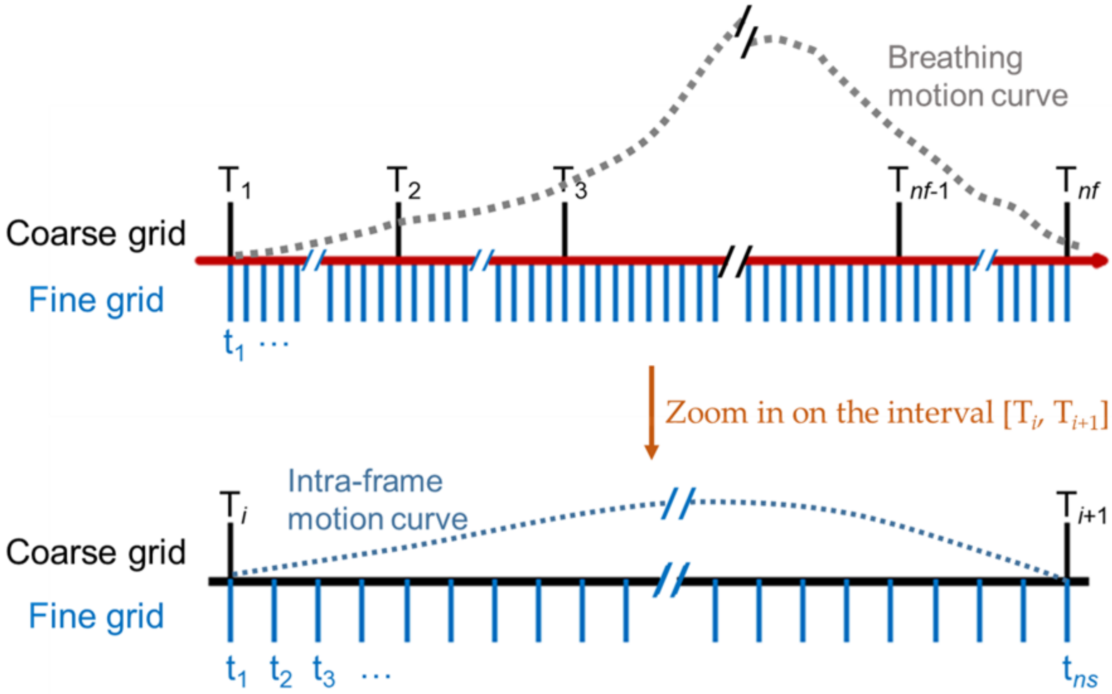

Figure 3.10: Coarse-to-fine grid scale representation of patient-specific motion data. The upper panel shows a coarse temporal grid (red) sampling key respiratory positions at time points T_1, T_2, ..., T_{nf}, aligned with the overall breathing motion curve (gray dotted line). Superimposed is a fine temporal grid (blue ticks) that densely samples within each coarse interval. The lower panel zooms in on one such interval [T_i, T_{i+1}] illustrating the intra-frame motion curve (blue dotted line) sampled by the fine grid [t_1, t_2, ..., t_{ns}].

In the first step (corresponding to Section 3.4.1), time-resolved volumetric MRI data are created, capturing key anatomical positions of human organs throughout the respiratory cycle. The breathing motion curve is discretized into nf phases, each corresponding to one of the nf frames in the sequence on the coarse grid [T_1, T_2, ..., T_{nf}].

In the second step (corresponding to Section 3.4.2), the intra-frame motion trajectory is depicted on a finer temporal grid, which subdivides the time intervals between consecutive coarse grid points into ns finer steps ([t_1, t_2, ..., t_{ns}]), corresponding to the ns temporal images. An intra-frame motion model, coupled with a motion pattern perturbation scheme, is introduced to enable a comprehensive representation of the real-world complexity, thoroughly exploring the potential anatomical variations during the frame acquisition period.

With the frames selected from the time-resolved volumetric MRI and the intra-frame motion data design method, it is possible to customize arbitrary synthetic yet realistic breathing motion curve, including a dedicated intra-frame motion trajectory. Subsequently, the simulation procedure introduced in Section 3.1 acts as the dataset generator, producing paired motion-corrupted images and ground-truth final-position images as needed.

3.4.1 4D MRI digital anthropomorphic phantom generation

3.4.1.1 Workflow

The 4D extended cardiac-torso (XCAT) phantom was developed to simulate realistic, highly detailed whole-body human anatomies for use in medical imaging research, encompassing thousands of anatomical structures. It incorporates parameterized models for both cardiac and respiratory motions, and provides users with significant flexibility to customize anatomical and motion variations [150]. In this section, the MRI version of the extended 4D XCAT phantom, referred to as the 4D MRI digital anthropomorphic phantom, is generated. The step-by-step workflow for this process is outlined in Fig. 3.11.

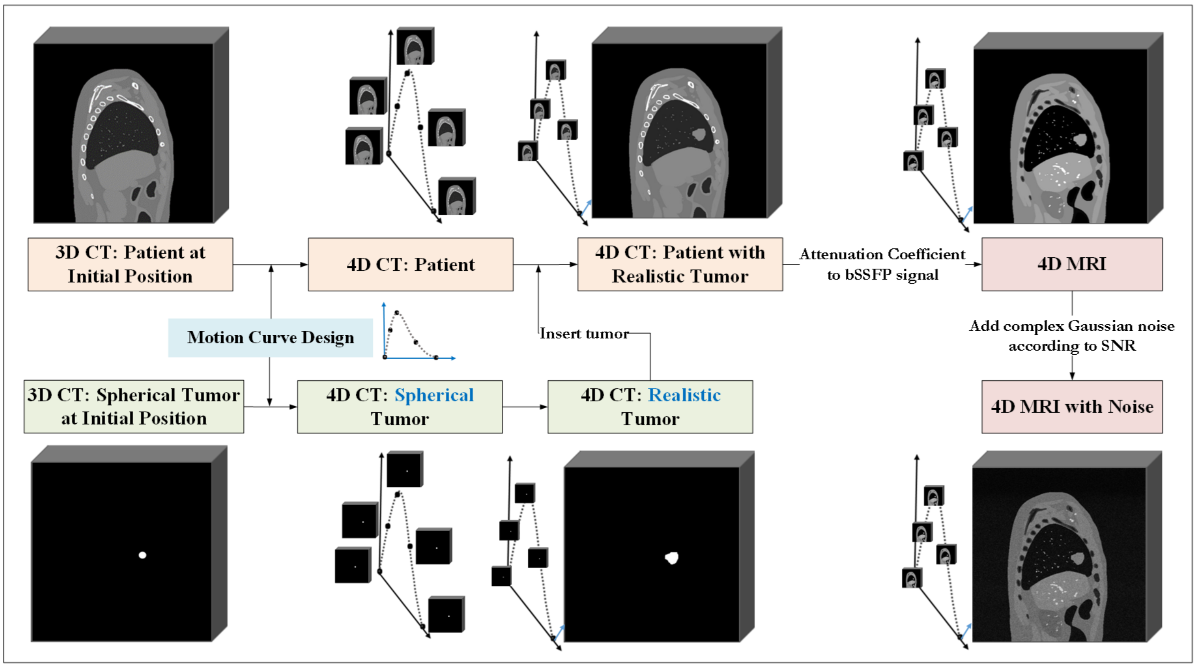

Figure 3.11: Workflow of 4D MRI digital anthropomorphic phantom generation. The phantom was schematically binned into 5 phases for each breathing cycle, but in the actual application, more breathing phases were used. This figure was originally published in [137].

Patient at initial position (3D CT) The workflow begins with a static representation of the virtual patient at the starting position of the breathing cycle, providing a baseline 3D CT volume of the anatomical structures. The complex shapes of real human organs are realistically modeled by setting detailed anatomical parameters in XCAT.

Spherical tumor at initial position (3D CT) Alongside the static virtual patient, a spherical tumor is positioned within a 3D CT volume, aligned with its initial position in the breathing cycle. The physical coordinates of the corresponding voxels for both the patient and tumor CTs are matched to ensure spatial consistency between the tumor and surrounding anatomy. This spherical tumor serves for localizing and propagating the centroid of a realistic tumor during motion, with the centroid—determined from its simplified geometric shape—acting as the reference for tumor motion tracking.

Motion Curve Design This step corresponds to defining motion on a coarse grid, as described in Fig. 3.10. In XCAT, respiratory motion is governed by two time-resolved curves: one indicating the variation in diaphragm height and the other describing the degree of chest expansion. This study defines these two curves by applying amplitude amplification coefficients (AAC) to a patient-specific respiratory motion waveform. Specifically, several types of motion waveforms are designed with amplitudes ranging from -10 to 0 mm (with negative values representing relative positions along the SI axis), indexed by frame number, to mimic both regular and irregular respiratory trajectories throughout the breathing cycle. Different AACs are then assigned to scale the waveform, characterizing the superior-inferior (SI) diaphragm motion and anterior-posterior (AP) chest-wall expansion. Additionally, tumor motions are categorized as either moving in sync with the surrounding lung tissues, or being guided by user-defined motion curves based on the waveform.

The beating heart motion in XCAT is defined by establishing parameters for the heart period, the timing of the cardiac cycle, and left ventricle volume at key phases: end-diastole, end-systole, the beginning of the quiet phase, the end of the quiet phase, and during reduced filling. The interaction between the cardiac and respiration motions is also accounted for [150]. By adjusting the translation or rotation parameters for heart respiratory motion, the extent of heart movements in specific directions during breathing can be tuned.

Patient with Realistic Tumor (4D CT) The motion curves are then applied to both the patient's anatomy and the spherical tumor, generating **Patient (4D CT)** and **Spherical Tumor (4D CT)**, respectively. Realistic tumors are initially segmented from treatment planning 4D CT scans of non-small cell lung cancer patients, relying on the exhale phase [151]. By aligning the centroid of a static 3D realistic tumor with the centroid positions extracted from the 4D CT of the spherical tumor, a 4D CT of the realistic moving tumor (**Realistic Tumor (4D CT)**) is obtained and subsequently integrated into the anatomical image (**Patient with Realistic Tumor (4D CT)**).

4D MRI Once the 4D CT phantom has been established, the anatomical data from the 4D CT is converted into 4D MRI data. This conversion is carried out by mapping the attenuation coefficient to the corresponding MRI signals of the same tissues. bSSFP pulse sequences, such as true fast imaging with steady-state precession (TrueFISP), which are typically performed for high-speed imaging, are of particular interest in this study for MRI signal simulation. The signal intensity in bSSFP ($\mathrm{S_{bSSFP}}$), with the RF pulses alternated by $180°$, is generally believed to be expressed as [89, 152]:

$$\mathrm{S_{bSSFP}} \propto \rho \sin\alpha \frac{1-\mathrm{e}^{-\mathrm{TR/T1}}}{1-\left(\mathrm{e}^{-\mathrm{TR/T1}}-\mathrm{e}^{-\mathrm{TR/T2}}\right)\cos\alpha-\left(\mathrm{e}^{-\mathrm{TR/T1}}\right)\left(\mathrm{e}^{-\mathrm{TR/T2}}\right)}\mathrm{e}^{-\mathrm{TE/T2}} \quad (3.13)$$

where α is the flip angle; T1, T2, and ρ are tissue-specific values for longitudinal relaxation, transverse relaxation, and proton density, respectively; To maintain signal stability and reduce the sensitivity of the sequence to magnetic field inhomogeneities, a very short TR interval (a few milliseconds) is used for bSSFP [153]. Therefore, TR $\ll$ T1 and TR $\ll$ T2, TR/T1 and TR/T2 approach 0. By evaluating the limit according to L'Hôpital's rule, Eq. 3.13 can be simplified as:

$$\mathrm{S_{bSSFP}} \propto \rho \sin\alpha \frac{1}{1+\cos\alpha+(1-\cos\alpha)(\mathrm{T1/T2})}\mathrm{e}^{-\mathrm{TE/T2}} \quad (3.14)$$

Tissue-specific parameters are determined following the reported values in the literature [89, 149, 151, 154], as summarized in Table 3.1.This study considers $\alpha = 60°$ and TE $= 1.27$ ms to match the acquisition parameters typically employed in the Viewray MRIdian [136] at LMU University Hospital. By converting the attenuation coefficient values in the 4D CT to the bSSFP signals for each tissue based on the corresponding T1, T2, and ρ maps, ideal noiseless 4D MRI phantoms are generated.

4D MRI with Noise To create more realistic MRI images, inherent noise present in real-world MRI acquisitions is simulated by adding independent and identically distributed (i.i.d.) complex Gaussian noise into the k-space $\mathrm{F}(k_x, k_y)$ of the noiseless 4D MRI, processed slice by slice. This results in additive Rician-distributed noise in the magnitude of the image domain:

$$\tilde{\mathbf{I}} = \mathcal{F}_2^{-1}\left(\mathrm{F}\left(k_x, k_y\right) + \delta_{Re} + j\delta_{Im}\right); \quad \delta_{Re}, \delta_{Im} \sim \mathcal{N}\left(0, \sigma^2\right) \quad (3.15)$$

where $\tilde{\mathbf{I}}$ denotes the noisy MR slices; $\mathcal{F}_2^{-1}$ represents the inverse 2D Fourier transform operator; and σ is the standard deviation of the Gaussian distribution, which can be derived from the predefined signal-to-noise ratio (SNR) using:

$$\sigma = \frac{\|\mathrm{F}\,(k_x, k_y)\|_2}{\sqrt{M}} \times \frac{10^{-\frac{\mathrm{SNR}}{20}}}{\sqrt{2}} \tag{3.16}$$

where M is the total number of elements of the k-space matrix.

The simulated time-resolved volumetric MRI phantoms effectively capture key anatomical positions throughout the breathing cycle, as represented in the coarse grid defined previously (Fig. 3.10). By altering the order of the frames in the sequences, it becomes possible to customize arbitrary complex breathing motion patterns. Therefore, 2D+t cine MR sequences can be obtained by extracting specific slices from these phantoms, enabling further investigation into the intra-frame motion of the target.

Table 3.1: Tissue-specific T1, T2, and ρ values used in calculating the bSSFP signal intensity. The values for ρ are reported in arbitrary units, relative to water. This table was originally published as supplementary material in [137].

	Background	Air lung /Bowel	Adipose	Water	Red marrow	Bowel	Pancreas	Muscle /Lesion	Kidney
T1 (ms)	0	0	376	376	276	122	909	825	921
T2 (ms)	0	0	30	30	13	8	28	28	40
ρ (a.u.)	0.00	0.00	1.00	1.00	0.32	0.09	0.85	2.39	1.48
	Heart	Liver	Spleen	Blood	Thyroid	Cartilage	Spine bone	Skull	Rib bone
T1 (ms)	1032	506	1466	1500	376	588	753	753	753
T2 (ms)	20	30	52	20	30	16	36	36	36
ρ (a.u.)	1.01	1.51	1.07	9.56	1.00	0.82	0.78	0.78	0.78

3.4.1.2 Basic information and motion data assignment for the simulated patients

According to the workflow outlined in the previous section (Section 3.4.1.1), a total of 25 4D MRI digital phantoms from lung cancer patients were generated, comprising 11 female, 11 male, and 3 adolescent subjects. Ten types of motion waveforms were designed using the amplitude-versus-frame-number curves, as shown in Fig. 3.12. In healthy adults at rest, the typical respiratory rate ranges from 12 to 15 breaths per minute, regulated by the respiratory center—typically involving an inhalation phase lasting approximately two seconds and an exhalation phase lasting around three seconds [155]. Considering the frame rate of cine-MR in currently commercially available MR-

Linac systems, the breathing cycle was binned in $nf = 20$ phases, approximating a 5-second breathing cycle period captured at around 4 FPS with cine-MR imaging.

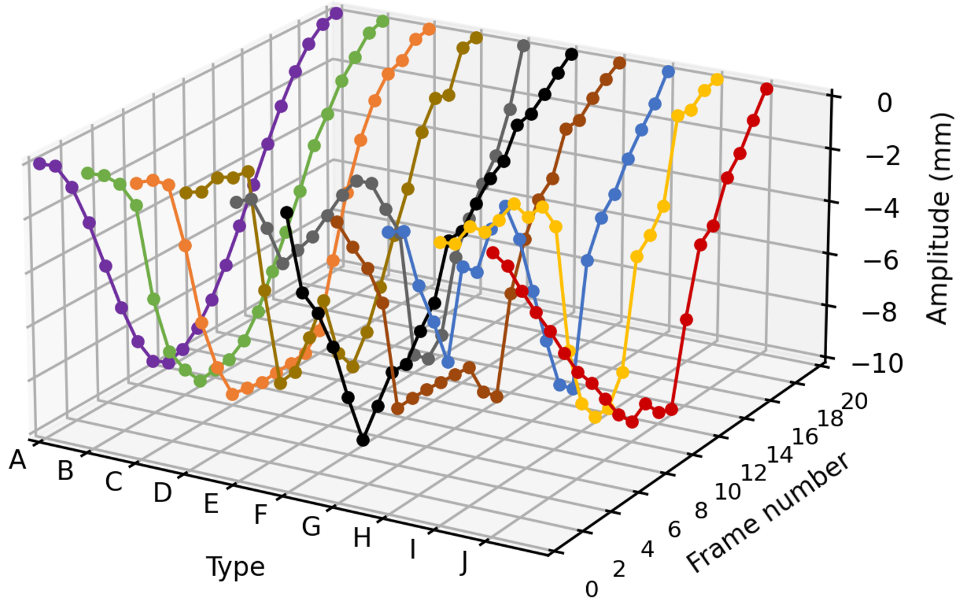

Figure 3.12: Ten types of designed patient-specific respiratory motion waveforms. These waveforms are scaled by amplitude amplification coefficients (AAC) to characterize the time-resolved motion of the diaphragm and chest wall. Intra-frame motion trajectories are excluded from this process.

Table 3.2 lists the basic information, assigned motion waveform types, and amplitude amplification coefficients for the simulated 25 patients. This study predominantly focused on fast motion, with most tumors located in the middle or lower lobes of the lung, where intra-frame motion is anticipated to be more pronounced. Based on published observations of respiratory motion in the investigated patients, the diaphragm can move up to 101 mm along the SI direction in a deep breathing mode [20]; the peak-to-peak lung tumor motion amplitude ranges $0 \sim 50$ mm in the SI direction and $0 \sim 24$ mm in the AP direction [20], while the maximum tumor speed is 72.6 ± 22.5 mm/s [99]. The motion parameter settings were designed based on these reported data, accounting for both typical and rapid movements. Table 3.3 summarizes the generated tumor motion data, including peak-to-peak and intra-frame motion values. Notably, Patient 10 exhibits the most significant tumor motion, with the largest peak-to-peak amplitude (56.4 mm) and the highest intra-frame displacement and speed (7.3 mm and 29.4 mm/s on average, respectively). It is important to note that these motion data represent only the position information of the anatomical structure at the exact beginning and end moment of each frame acquisition in the original sequence. A more detailed information of intra-frame motion trajectories will be presented in the following sections.

The length of the heart beating cycle was set to 1.0 second for all the patients. Specifically, the duration from end-diastole to end-systole was defined as 0.5 seconds, from end-systole to beginning of quiet phase was 0.192 seconds, the quiet phase lasted 0.115 seconds, and from end of quiet phase to reduced filling was 0.193 seconds. The cardiac motion during respiration was modeled as a rigid translation of 0.5 cm in the AP direction and 2 cm in the SI direction, with no rotational component.

Fig. 3.13 presents examples of cine-MR frames obtained from the simulated patients (Patient 15 and Patient 17). The left panel displays the end-diastole and end-systole cardiac phases, extracted from the breath-hold process, as depicted in the Type C waveform (see Fig. 3.12). The right panel illustrates the end-expiration and end-inspiration phases, during which the tumor exhibits motion in synchrony with lung deformation. Due to out-of-plane displacement, the tumor is not visible in the current axial and coronal slices at the end-inspiration phase.

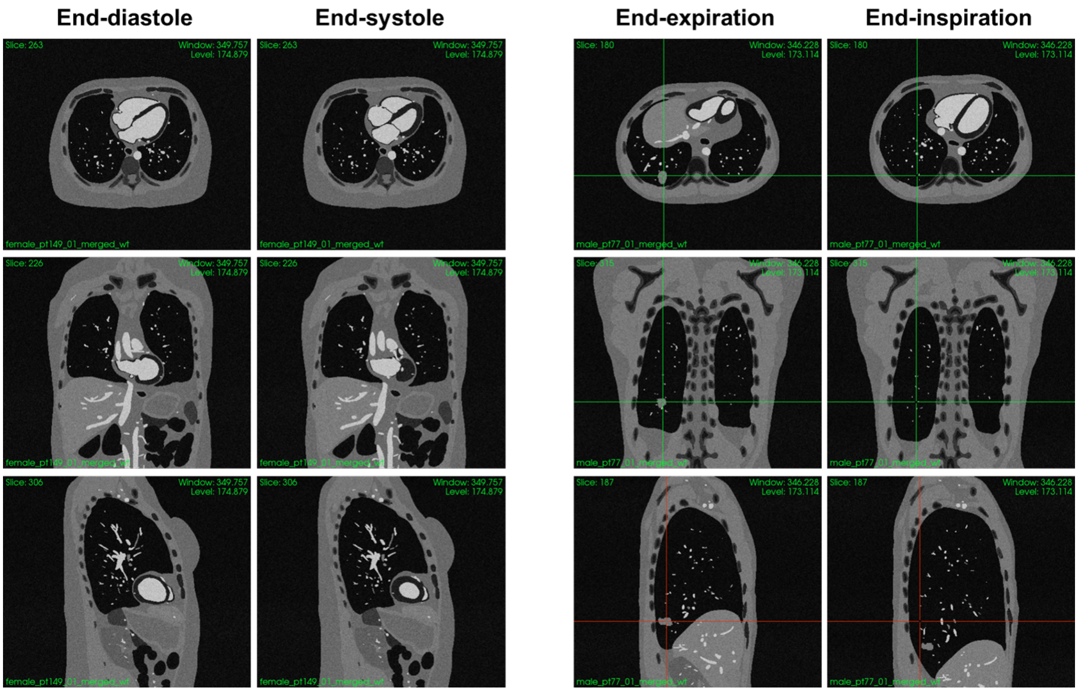

Figure 3.13: Examples of simulated cine-MR frames from Patient 15 (left) and Patient 17 (right). The selected cardiac and respiratory phases include end-diastole, end-systole, end-expiration, and end-inspiration, presented in axial (top), coronal (middle) and sagittal (bottom) views. In the right panel, the reference lines intersect at the tumor in the end-expiration positions.

Table 3.2: Basic information and the breathing motion curve assignment for the simulated patients. Tumor location in the lung is presented as R-Right, L-Left/ l-lower, m-middle, u-upper(lobe) / P-Posterior, A-Anterior, M-Middle; AAC indicates amplitude amplification coefficient.

Patient ID	Gender	Age	Weight (kg)	Height (cm)	BMI	Tumor location	Waveform type	Diaphragm AAC	Chest-wall AAC
01	F	63	81.3	153	34.73	R/l/P	A	2	-1.2
02	F	65	78.6	161	30.32	L/l/P	B	3	-1.1
03	F	57	105.8	165.1	38.81	R/m/M	E	4	-1.3
04	F	65	56	164.7	20.64	L/l/A	D	3	-1.2
05	F	56	69.6	166.76	25.03	R/m/M	B	5	-1
06	M	63	72.1	170	24.95	L/l/M	A	4	-1.6
07	M	70	100.4	173.7	33.28	R/l/A	E	2	-0.9
08	M	52	60.75	173	20.30	L/l/P	D	4	-1.2
09	M	67	89.9	178.5	28.22	R/m/A	C	5	-1.4
10	M	50	120	177.8	37.96	L/l/P	F	6	-1.5
11	F	27	55.6	172.7	18.64	R/l/A	H	4	-1.9
12	F	37	78.7	169.5	27.39	R/m/P	G	3	-1.2
13	F	49	105.1	172	35.53	L/u/M	J	5	-1
14	F	51	68.2	175	22.27	L/l/A	I	3	-0.9
15	F	40	75.4	160	29.45	R/l/P	C	2	-1
16	F	52	86	153	36.74	R/m/P	F	4	-1.3
17	M	31	77.9	185.2	22.71	R/m/P	J	6	-2
18	M	58	117	180	36.11	L/l/A	H	4	-1.3
19	M	18	62	176	20.02	L/l/A	I	5	-1.7
20	M	63	75.6	167.7	26.88	R/l/P	G	6	-1.8
21	M	64	84.15	180	25.97	L/u/A	A	4	-1.1
22	M	60	88	190	24.38	L/l/A	J	5	-1.8
23	F	16	59.9	173.5	19.90	L/l/A	F	3	-0.9
24	M	14	67.4	181.06	20.56	R/m/M	E	2	-0.6
25	F	11	31.1	135.1	17.04	R/u/P	C	3	-0.8

Table 3.3: Tumor motion characteristics for the simulated patients, detailing both peak-to-peak motion amplitudes and intra-frame motion. Intra-frame motion displacement and average speed values are presented as mean [max] over all 20 frames, covering a full breathing cycle. The patient exhibiting the most significant tumor motion is shown in bold.

Patient ID	Peak-to-peak motion			Intra-frame motion			
	Amplitude (mm)			Displacement (mm)			Avg. speed (mm/s)
	SI	AP	Total	SI	AP	Total	
01	15.8	9.5	18.5	1.6 [2.9]	1.0 [1.7]	1.9 [3.4]	7.6 [13.7]
02	23.9	8.8	25.5	2.5 [6.3]	0.9 [2.3]	2.6 [6.7]	10.5 [26.8]
03	27.1	7.6	28.1	2.7 [8.1]	0.8 [2.0]	2.8 [8.3]	11.3 [33.3]
04	24.9	10.6	27.0	2.7 [8.6]	1.1 [3.7]	2.9 [9.3]	11.7 [37.3]
05	29.8	6.0	30.4	3.0 [7.8]	0.6 [1.6]	3.1 [8.0]	12.4 [31.8]
06	37.7	12.7	39.8	3.9 [7.2]	1.3 [2.5]	4.1 [7.6]	16.5 [30.3]
07	17.4	8.0	19.2	1.8 [5.3]	0.8 [2.3]	1.9 [5.8]	7.8 [23.1]
08	31.2	9.4	32.5	3.5 [10.9]	1.0 [3.2]	3.6 [11.3]	14.5 [45.3]
09	41.5	13.2	43.5	4.3 [10.1]	1.4 [3.2]	4.5 [10.6]	17.9 [42.2]
10	**55.5**	**10.1**	**56.4**	**7.2 [19.0]**	**1.3 [3.5]**	**7.3 [19.3]**	**29.4 [77.1]**
11	33.3	16.0	36.9	4.5 [11.5]	2.2 [5.5]	5.0 [12.8]	20.1 [51.1]
12	23.3	7.7	24.5	2.4 [7.5]	0.8 [2.5]	2.5 [7.9]	10.1 [31.5]
13	36.8	8.0	37.7	3.7 [9.3]	0.8 [2.0]	3.7 [9.5]	15.0 [38.0]
14	26.0	8.0	27.2	2.7 [10.1]	0.8 [3.1]	2.9 [10.6]	11.5 [42.4]
15	17.0	7.3	18.5	1.8 [4.7]	0.8 [2.0]	1.9 [5.1]	7.7 [20.6]
16	31.2	9.3	32.6	3.2 [10.7]	0.9 [3.2]	3.3 [11.2]	13.3 [44.6]
17	49.4	16.9	52.2	4.9 [12.5]	1.7 [4.3]	5.2 [13.2]	20.7 [52.7]
18	34.0	10.7	35.7	4.6 [11.8]	1.5 [3.7]	4.9 [12.3]	19.4 [49.4]
19	44.8	14.6	47.1	4.7 [17.5]	1.5 [5.7]	5.0 [18.4]	19.9 [73.7]
20	52.8	14.1	54.7	5.4 [17.0]	1.5 [4.5]	5.6 [17.5]	22.4 [70.2]
21	28.1	9.7	29.8	2.9 [5.3]	1.0 [1.8]	3.1 [5.6]	12.3 [22.4]
22	46.5	16.9	49.5	4.6 [11.7]	1.7 [4.3]	4.9 [12.5]	19.6 [49.9]
23	26.6	7.6	27.6	2.7 [9.1]	0.8 [2.6]	2.8 [9.5]	11.3 [37.9]
24	16.9	5.2	17.6	1.8 [6.2]	0.5 [1.9]	1.9 [6.5]	7.4 [26.1]
25	21.5	5.7	22.3	2.2 [6.0]	0.6 [1.6]	2.3 [6.2]	9.2 [25.0]
Avg.	31.7	10.1	33.4	3.4 [9.5]	1.1 [3.0]	3.6 [10.0]	14.4 [39.9]

3.4.2 Intra-frame motion data

In this section, intra-frame motion data is represented as displacement vector fields. To design the DVFs and comprehensively capture the coverage of potential trajectories across the full range of anatomical positions, an intra-frame motion model and a dedicated motion pattern perturbation scheme are proposed.

3.4.2.1 Intra-frame motion model

The intra-frame motion model is constructed with a piecewise linear approximation between consecutive control points. Specifically, the overall frame acquisition time step interval, $[1, ns]$, is subdivided into multiple consecutive intervals, with the endpoints referred to as control points. Motion between control points is represented by DVFs derived from corresponding images and subsequently discretized over time steps. The optical flow-based DIR algorithm [85] is employed to estimate the DVFs.

To minimize errors introduced by optical flow and obtain the most accurate possible final-position image as ground truth, intra-frame motion data (DVF_m) and temporal images $\mathbf{I_j}$ ($j = i, i+1, \ldots, i+m$) for a specific sub-interval $[i, i+m]$ within $[1, ns]$ are determined as follows:

$$\begin{aligned} \text{DVF}_m &= \arg\min \text{MSE}\left(\mathbf{I_{i+m}} \oplus \text{dvf},\ \mathbf{I_i}\right), \\ &\text{where} \quad \text{dvf} \in \{\text{DVF}_{i+m\to i},\ -\text{DVF}_{i\to i+m}\}\ ; \\ \mathbf{I_j} &= \mathbf{I_{i+m}} \oplus \left(\tfrac{i+m-j}{m} \times \text{DVF}_m\right), \quad j = i,\ i+1,\ \ldots,\ i+m. \end{aligned} \tag{3.17}$$

where the symbol $\oplus$ denotes the image deformation based on the given DVF; MSE$(,)$ refers to the mean squared error (MSE) computation between two images; $\mathbf{I_{i+m}}$ is the image at control point $i+m$, and $\mathbf{I_i}$ is the image at control point i; $\text{DVF}_{i+m\to i}$ represents the DVF from $\mathbf{I_{i+m}}$ to $\mathbf{I_i}$, while $\text{DVF}_{i\to i+m}$ represents the DVF from $\mathbf{I_i}$ to $\mathbf{I_{i+m}}$. Theoretically, $\text{DVF}_{i+m\to i}$ and $-\text{DVF}_{i\to i+m}$ should be identical; however, due to limitations in the accuracy of the optical flow algorithm, the DVF yielding the lower residual MSE (i.e., a better $\mathbf{I_i}$ restoration) after registration is selected.

3.4.2.2 Intra-frame motion pattern perturbation scheme

Once the 2D+t cine MR sequences have been selected from the 4D MRI data (detailed in Section 3.4.1), nf key anatomical positions throughout each breathing cycle are identified. An intra-frame motion pattern perturbation scheme is then introduced to determine the images at the control points, as discussed in Section 3.4.2.1.

First, nf key-frame sets and the corresponding images are defined and labeled from the original 2D+t cine MR sequences. To achieve this, four additional frames are interpolated between two consecutive frames in the original sequence. Fig. 3.14 presents a schematic view of this step: the k-th frame in the original sequence ($k = 1, 2, \ldots, nf$) is labeled as k^1 (black dot in the figure). Images at k^2, k^3, k^4, and k^5 are generated based on linear interpolation of the DVFs between the corresponding frames of k^1 and $(k+1)^1$. The five images are thus considered to fall within the k-th key-frame set, comprising k^1 from the original sequence and four interpolated frames k^2, k^3, k^4, and k^5. This process can also be seen as an efficient way to increase the temporal resolution of the original cine-MR sequence by a factor 5, avoiding the significant time required to directly generate 4D MRI phantoms with 5× temporal resolution. It effectively enhances the diversity of anatomical positions for the ground truth and introduces randomness within a specified range for each control point in the following step.

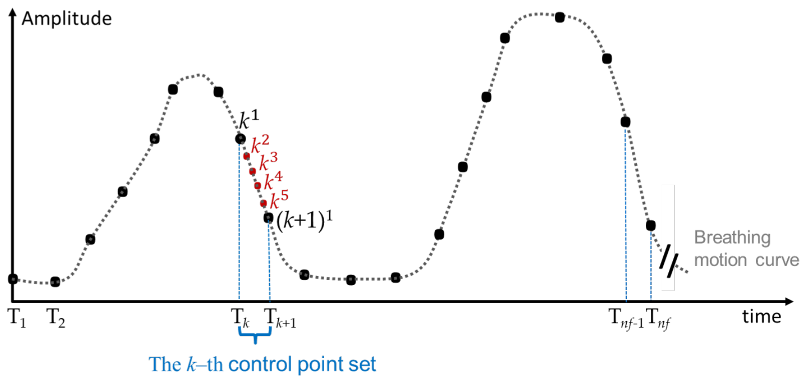

Figure 3.14: Schematic illustration of the definition of the k-th key-frame set on the original 2D+t cine MR sequences.

Next, intra-frame motion trajectories are manipulated by varying the number or order of control point images: first, the number of control points governing the intra-frame motion trajectory is specified; then for each control point, one of nf key-frame sets is assigned, followed by randomly selecting one image from the chosen set of five as the control point image. The overall motion extent can be controlled by adjusting the key-frame set indices for consecutive control points, based on their positions in the original sequence.

Consequently, the original intra-frame motion pattern of the cine-MR sequence, utilizing linear DVF decomposition between consecutive frames in relation to the time step, is expanded to include a variety of patterns as required. Table 3.4 lists the configurations of the proposed motion patterns, several of which will be applied in subsequent chapters to create datasets that support intra-frame motion compensation in Cartesian and radial cine-MRI. The proposed patterns incorporate two, three, or four control points to progressively enhance the degrees of freedom, thereby accommodating increased motion irregularity.

Table 3.4: Configurations of the designed intra-frame motion patterns. The letter "S" indicates a sudden application of the rigid motion, while "L" denotes a linear application.

Pattern	Number of control points	Key-frame set index			Apply rigid motion	Identical control point images
		First control point	Middle control point(s)	Last control point		
01	2	k	–	k	No	Yes
02	2	k	–	$k+1$	No	No
03	2	k	–	$k-1$	No	No
04	3	k	k	$k+1$	No	No
05	3	k	$k+1$	$k+1$	No	No
06	3	k	$k+1$	$k-1$	No	No
07	3	k	$k-1$	k	No	No
08	3	k	$k-2$	k	No	No
09	3	k	$k+1$	$k+2$	No	No
10	3	k	$k-2$	$k-2$	No	No
11	3	k	$k+2$	$k+4$	No	No
12	3	k	$k+4$	$k+3$	No	No
13	3	k	$k-3$	$k-5$	No	No
14	3	k	$k-1$	$k+1$	Yes / S	No
15	3	k	$k+2$	k	Yes / S	No
16	3	k	k	$k+1$	Yes / L	No
17	3	k	$k-1$	k	Yes / L	No
18	3	k	k	k	Yes / S	No
19	4	k	$k+1$, k	$k-2$	No	No
20	4	k	$k-1$, $k+1$	$k+2$	No	No

Motion patterns with two control points adopt $i = 1$ and $m = ns - 1$ in Eq. 3.17, with Pattern 2 corresponding to the original intra-frame motion pattern. It is essential to emphasize that, in the case of a static scenario (Pattern 1), the control point images remain identical to ensure the absence of intra-frame motion. In this context, the output image produced by the compensation model is anticipated to be the same as the input.

For patterns with three or four control points, random insertion moments are chosen for the middle control points, effectively dividing $[1, ns]$ into two or three sub-intervals of random lengths. To simulate an overall target drift during the frame acquisition, an additional rigid motion is applied in the second sub-interval (denoted as $[mp, ns]$, with the middle control point represented as mp) in specific cases involving three control points (Pattern 14 $\sim$ 18). The parameters for the rigid transformation are determined by selecting a random value for the rotation angle within the range $[-\pi/20, \pi/20]$, and a translation extent along each axis within the range $[-1, 1]$ pixels.

Two methods are considered for applying the rigid motion: a sudden application and a linear application. Let $\mathbf{I}^R$ represent the image obtained after applying a rigid transformation to image $\mathbf{I}$. In the case of a sudden application, the control point images for the sub-interval $[mp, ns]$ are $\mathbf{I}^R_{mp}$ and $\mathbf{I}^R_{ns}$; whereas, in the case of a linear application, they are $\mathbf{I}_{mp}$ and $\mathbf{I}^R_{ns}$.

In summary, three degrees of freedom are incorporated in the motion pattern perturbation scheme: randomly selection of images from the key-frame sets, the insertion moments of the middle control points, and the rigid motion parameters. These elements introduce randomness into the database, allowing a comprehensive exploration within the domain of potential anatomical structure positions. Some extreme scenarios, which may never occur in reality, are crafted to create lager differences between motion-corrupted and ground-truth final-position images, compelling the potential network to focus more on the dynamic mechanisms and remain robust against variations in motion amplitude.

Using the determined control point images as inputs, intra-frame motion data and corresponding temporal images are generated leveraging the motion model expressed in Eq. 3.17. The simulation procedure introduced in Section 3.1 functions as the dataset generator, effectively producing input-output training pairs (intra-frame-motion-corrupted images and ground-truth final-position images corresponding to the last shot of the frame acquisition) as required for the labeled dataset.

3.4.3 Examples of motion-corrupted images

This section presents several examples of generated motion-corrupted images for Patient 02 and Patient 08, as they moved from the initial position to the final position during frame acquisition, with corresponding images shown in Fig. 3.15. Patient 02 was inhaling throughout the k-space sampling, causing the tumor to shift generally downwards, while Patient 08 was exhaling, resulting in an upward tumor movement.

In Fig. 3.15, motion-corrupted images were simulated following motion Pattern 02, with a *linear* Cartesian phase encoding direction orthogonal to the main direction of intra-frame motion. Reference lines mark the upper and lower boundaries of the tumors' ground-truth position at the conclusion of the frame acquisition.

It is evident that the tumor positions derived from motion-corrupted images lagged behind the actual final positions, clearly indicating noticeable imaging latency. Quantitatively, the latency was approximately 50% of the frame acquisition time, consistent with the conclusions discussed in Section 3.1.2.2. Compared to target positioning errors, the impact of motion artifacts (or image blur) was negligible in the overall imaging errors. In Fig. 3.15, the anatomical geometry remained well-preserved in the motion-corrupted images. This observation differs from imaging systems acquiring signals directly in the image domain, such as fluoroscopy, where the detector may capture the target's entire path (passing pixels) during acquisition.

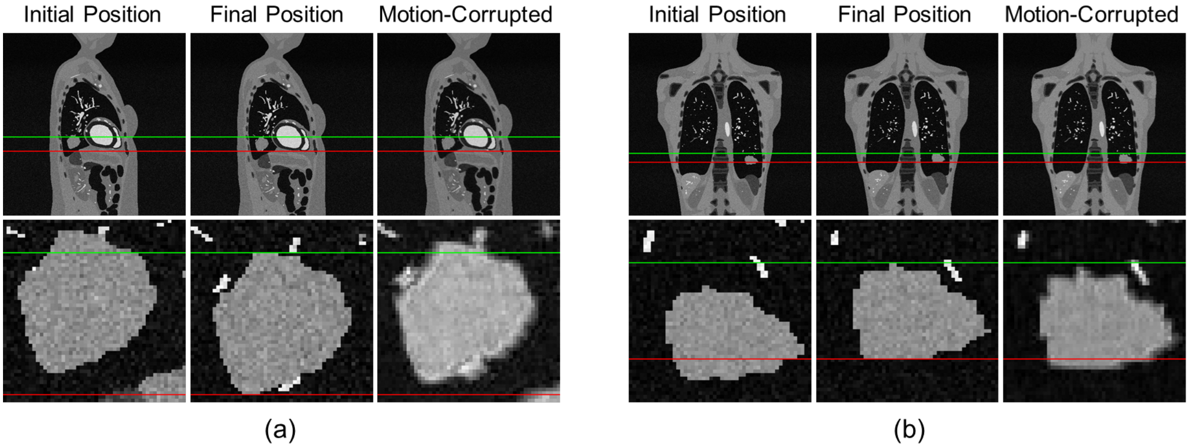

Figure 3.15: Examples of generated motion-corrupted images for (a) Patient 02 and (b) Patient 08. Each panel displays the patient's progression from the initial position (left) to the final position (middle) according to motion Pattern 02. The resulting motion-corrupted image is shown in the right column. Enlarged views of the tumor, captured at identical coordinates, are provided with reference lines marking the upper and lower boundaries of the ground-truth final position. The $linear$ Cartesian phase encoding direction is anterior-posterior (AP) in (a), and left-right (LR) in (b).

In current clinical practice with MR-Linac, online anatomy tracking or beam gating is achieved based on target deformation using DVFs, estimated through deformable image registration from a reference frame to live cine-MR frames [30]. Fig. 3.16 demonstrates visually the errors in DVF determination caused by cine-MR intra-frame motion. The selected initial- and final- position frames ($\mathbf{I}_1$ and $\mathbf{I}_{ns}$) were the same as those in Fig. 3.15. The motion-corrupted image $\mathbf{I}_{\mathbf{motion}}$ was also simulated according to motion Pattern 02 from $\mathbf{I}_1$ to $\mathbf{I}_{ns}$, with different k-space phase encoding directions being considered. The measured DVF, derived from $\mathbf{I}_{\mathbf{motion}}$, was compared to the ground truth, which was obtained from $\mathbf{I}_{ns}$. To facilitate an intuitive comparison, the reference frame $\mathbf{I}_{\mathbf{ref}}$ for image registration was specifically selected as the ground-truth final-position image: $\mathbf{I}_{\mathbf{ref}} = \mathbf{I}_{ns}$. Under this condition, the ground-truth DVF was set to $\mathbf{0}$, and the DVF from $\mathbf{I}_1$ to $\mathbf{I}_{ns}$, which reflects intra-frame motion, should have the same magnitude as the DVF from $\mathbf{I}_{\mathbf{ref}}$ to $\mathbf{I}_1$ but in the opposite direction.

The results indicated residual intra-frame motion components in the measured DVF, highlighting substantial errors in DVF determination due to intra-frame motion deterioration effects. The dominant component of the intra-frame anatomical changes occurred along the SI direction. Qualitatively, compared to an orthogonal phase encoding direction, slightly greater errors were appreciable when phase encoding was applied in the SI direction.

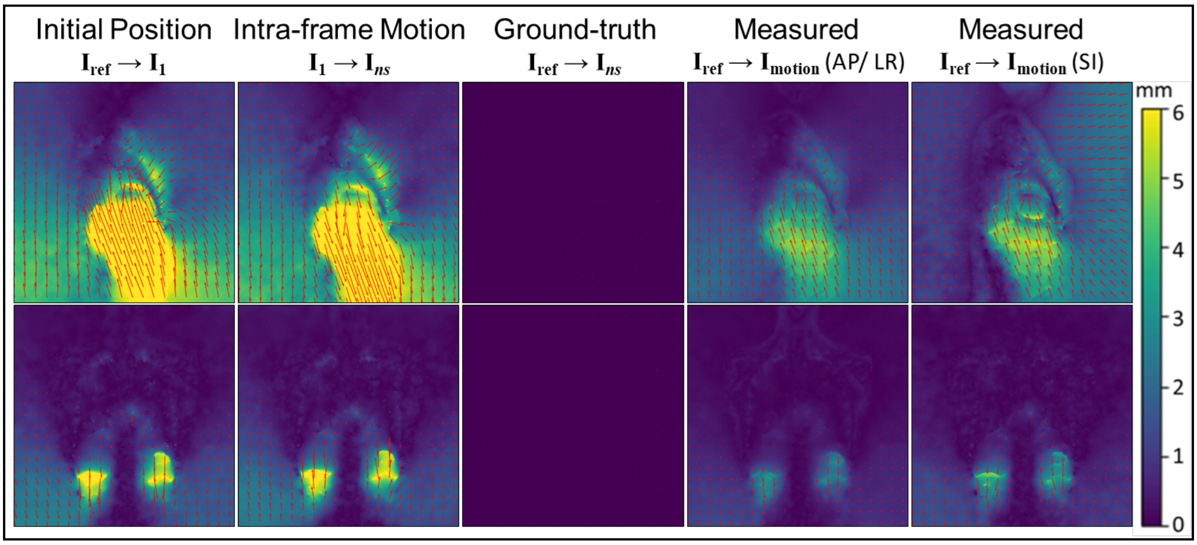

Figure 3.16: Displacement vector fields for Patient 02 (top) and Patient 08 (bottom). From left to right: DVF from the reference frame to initial-position image ($\mathbf{I}_{\text{ref}} \rightarrow \mathbf{I}_1$); DVF of the intra-frame motion ($\mathbf{I}_1 \rightarrow \mathbf{I}_{ns}$); Ground-truth DVF ($\mathbf{I}_{\text{ref}} \rightarrow \mathbf{I}_{ns}$); Measured DVF ($\mathbf{I}_{\text{ref}} \rightarrow \mathbf{I}_{\text{motion}}$), with $linear$ Cartesian phase encoding direction either orthogonal (AP/LR) or parallel (SI) to the main direction of intra-frame motion. This figure is adapted from material originally published in [137].

Fig. 3.17 shows examples of motion-related imaging errors across various motion patterns and phase encoding directions in a $linear$ Cartesian trajectory, with the selected initial and final positions consistent with those in Fig. 3.15. The results demonstrate that intra-frame motion patterns significantly impact the extent of image degradation, resulting in variations in anatomy tracking accuracy. Specifically, an insertion moment at 65% of the acquisition time (mp = 65% ns) in motion pattern 04 ($\mathbf{I}_1 \rightarrow \mathbf{I}_1 \rightarrow \mathbf{I}_{ns}$) led to poorer image quality. This is consistent with expectations, as the lower frequency components of k-space, which are of a much higher magnitude and primarily determine the target position in Cartesian readout trajectories—originate predominantly from the initial-position image in this scenario. Similar to the conclusions drawn from the DVF analysis, the choice of phase encoding direction qualitatively has a minor effect on contouring accuracy, with the SI direction exhibiting slightly more motion artifacts compared to the other direction. Nevertheless, the contribution of image blur to the overall imaging errors is negligible when considering the more significant factor of the target positioning errors.

For radial sampling, Fig. 3.18 presents examples of motion-induced imaging errors from Patient 02, simulated using the same initial and final position images as in Fig.

3.15. This figure compares the effects of applying various motion patterns (Pattern 02 and Pattern 04) and k-space readout trajectories, including *linear* (ψ_{linear}), *golden angle* (ψ_{gold}) and *tiny golden angle* (ψ_5, ψ_{10}). In each trajectory, the starting angle of the first spoke, γ, was set to a random value. The results reveal negligible variations in anatomy positioning accuracy across different radial trajectories, though slight artifacts are perceptible in the *linear* case. Motion Pattern 04 resulted in larger imaging errors than Pattern 02, but the difference between them is relatively small in comparison to those presented in Fig. 3.17. The findings indicate a uniform contribution from each spoke to the reconstruction of the target position in the presence of intra-frame motion, regardless of its spatial orientation or distribution within the radial trajectory.

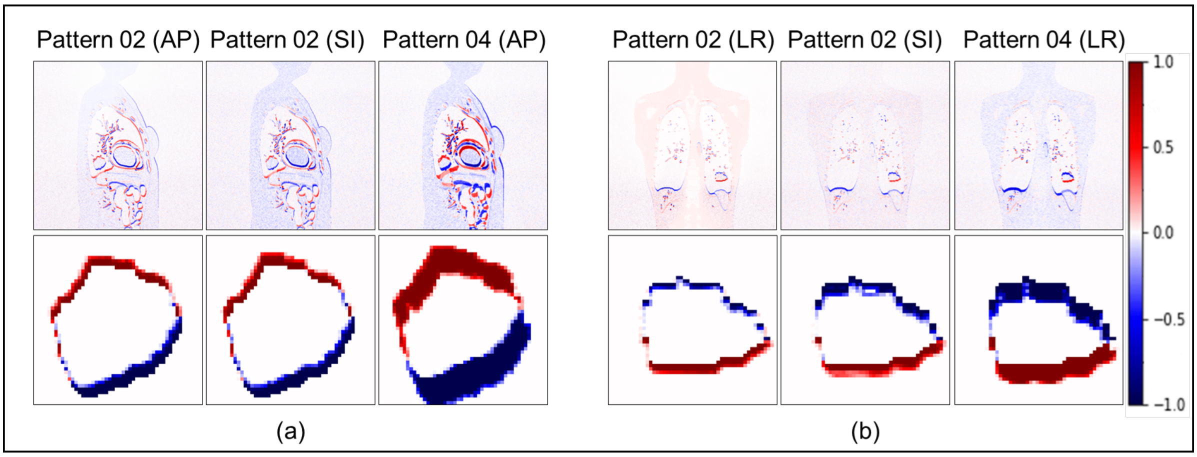

Figure 3.17: Examples of motion-related imaging errors resulting from various motion patterns and phase encoding directions in a *linear* Cartesian trajectory. Displayed are difference images and tumor contouring errors between the training pairs, specifically motion-corrupted images and their corresponding ground-truth final-position images, for (a) Patient 02 and (b) Patient 08. The difference values are calculated as the motion-corrupted minus the ground-truth. Phase encoding directions are indicated in brackets, including anterior-posterior (AP), superior-inferior (SI) and left-right (LR). For motion Pattern 04, the middle-point insertion moment occurs at 65% of the acquisition time (mp = 65% ns). This figure is adapted from material originally published in [137].

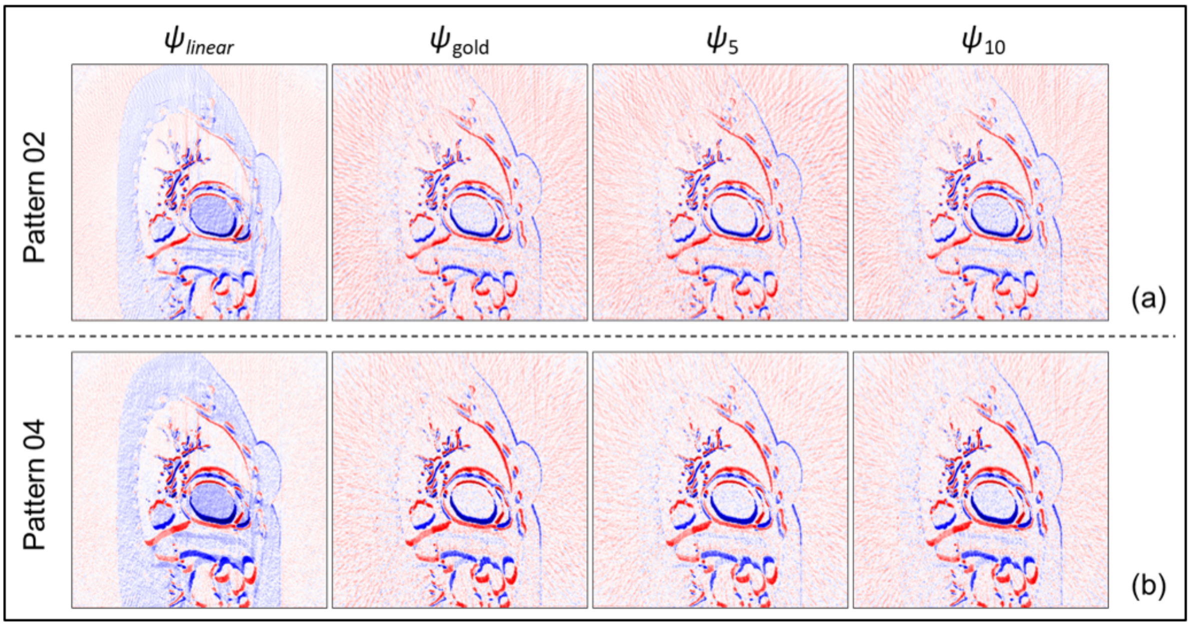

Figure 3.18: Examples of imaging errors with (a) Motion Pattern 02 and (b) Motion Pattern 04, under varying azimuthal profile increments in radial k-space sampling trajectories. From left to right: ψ_{linear}, ψ_{gold}, ψ_5, ψ_{10}. The starting angle of the first spoke, γ, was set randomly in each trajectory. Displayed are difference images between training pairs from Patient 02, specifically motion-corrupted images and their corresponding ground-truth final-position images. The difference values are calculated as the motion-corrupted minus the ground-truth. For motion Pattern 04, the middle-point insertion moment occurs at 65% of the acquisition time (mp = 65% ns).

Chapter 4

INTRA-FRAME MOTION COMPENSATION FOR CARTESIAN CINE-MRI

4.1 Method and materials

4.1.1 Model

The selection of the network architecture should account for the specific characteristics of the inverse problem that need to be addressed. Fig. 4.1 shows an example of a motion-corrupted image decomposition experiment for Cartesian sampling. The image was simulated according to Motion Pattern 18 (see Section 3.4.2.2), with *linear* phase encoding applied along the AP direction. Under these conditions, the later-acquired data correspond to the higher frequency components on the right-hand side of the Fourier domain. A sudden 9° rotation was introduced at the middle-point insertion moment, occurring at 70% of the acquisition time.

In the figure, the motion-corrupted image retains the same anatomical position as the initial location, which is expected to be corrected to align with its corresponding final-position image by the compensation model. The decomposition of the motion-corrupted image reveals that the final-position contour is encoded in the motion-corrupted k-space. However, due to the orders-of-magnitude difference in values between the low- and high-frequency components, these true-position details are obscured by the dominant lower-frequency information, making them difficult to discern visually in the spatial domain.

Therefore, the intuitive concept of an intra-frame motion compensation model is to detect and extract information from the later-acquired data, which can subsequently guide the processing of the earlier-acquired components. For a *linear* Cartesian sampling trajectory, this process is akin to determining the final-position contour—often imperceptible from the motion-corrupted image—and filling the contour with the corrected LFC-associated patterns.

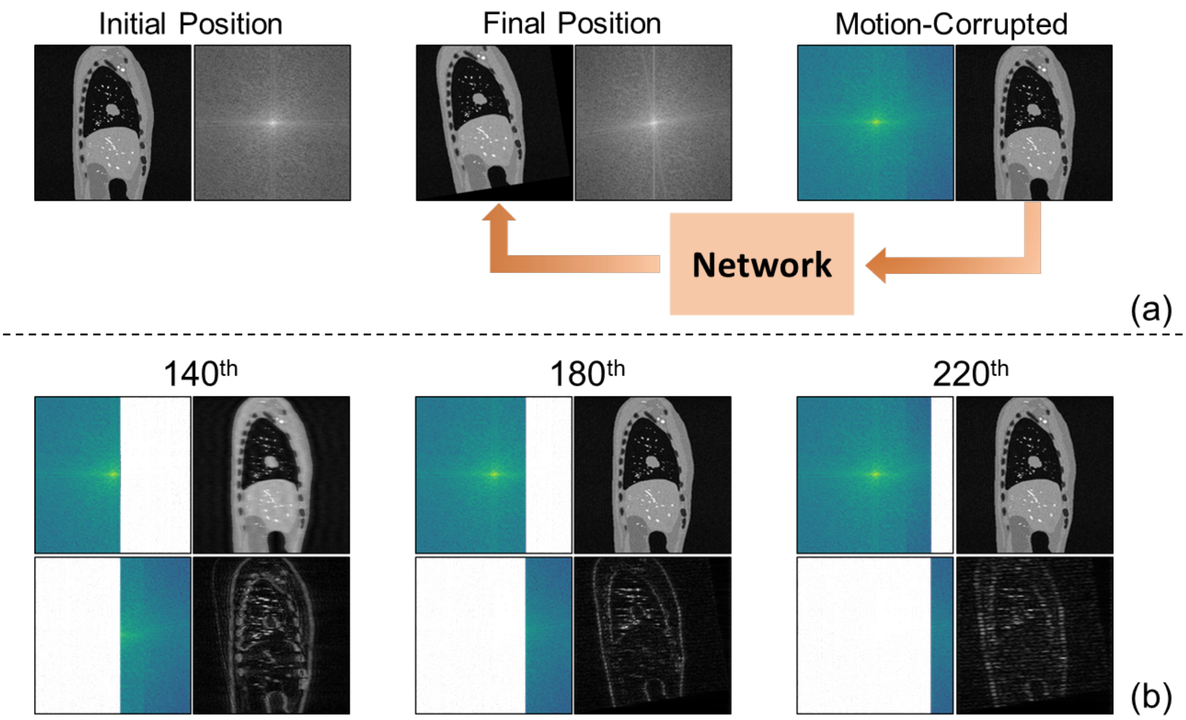

Figure 4.1: Motion-corrupted image decomposition experiment for *linear* Cartesian sampling. (a) The motion-corrupted image and corresponding k-space, simulated according to Motion Pattern 18, with the phase encoding direction along the AP direction. A sudden 9° rotation is introduced at the middle-control-point insertion moment, which occurs at 70% of the acquisition time. Initial and final position images and their corresponding k-space are shown on the left. The matrix size of the images is 256 $\times$ 256. (b) Decomposition of the motion-corrupted k-space/image at three specific temporal positions (220th, 180th, and 140th time steps), displaying the preserved frequency components (left) and the corresponding reconstructed images (right). The blank regions in the k-space represent zero-valued areas. The images are normalized to the range [0, 1], and the k-space is represented on a logarithmic scale.

The properties of convolutional neural networks make them promising models to fulfill these requirements. By utilizing a series of building blocks such as convolutional layers, pooling layers, and fully connected layers, CNNs are structured to automatically and adaptively learn spatial hierarchies of features through backpropagation [156], making them powerful models for feature extraction in pattern recognition, semantic image segmentation, and various other tasks. Recently, CNN functionality has become more interpretable through explanation techniques involving frequency component decomposition [157]. Wang et al. observed CNNs' ability to capture HFCs in images, which are largely indiscernible to human perception [158].

The architecture of CNN models can be highly flexible. In the context of intra-frame motion compensation in *linear* Cartesian k-space trajectories, later-acquired data correspond to the HFCs and must be preserved, while the patterns associated with the LFCs are processed. Therefore, the UNet architecture [159], initially designed for biomedical image segmentation, was employed to enable end-to-end training for directly deriving the final-position image from the motion-corrupted input. The concatenative skip connections in UNet transfer features from encoder to decoder at the same dimensionality, supporting the recovery of fine-grained details lost during down-sampling.

Fig. 4.2 shows the typical 5-level UNet architecture exploited in this study. The real and imaginary parts of the input and output images are represented as separate channels. Each level of the network comprises a double convolution block using 3×3 convolution kernels, followed by batch normalization and ReLU activation. The first level has 64 feature channels, $n_{\text{ch1}} = 64$, which are then sequentially doubled in the subsequent levels. A 2×2 max pooling operation with stride 2 is applied for down-sampling in the contracting path, while "up-convolution" (also referred to as transposed convolution) is implemented for up-sampling in the expansive path followed by concatenation. A 1×1 convolutional layer is set as the final layer of the network, which ultimately provided the output image.

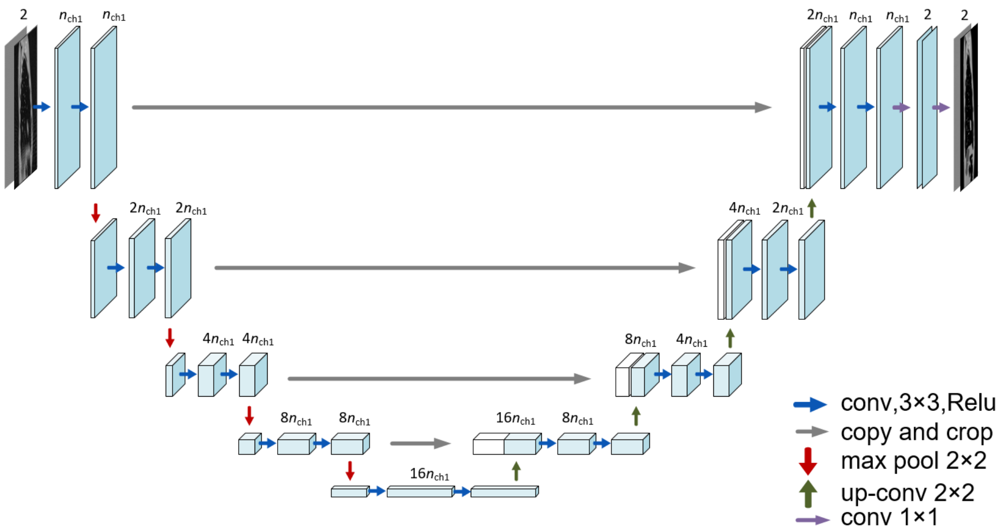

Figure 4.2: UNet architecture: Blue boxes represent multi-channel feature maps, while white boxes indicate copied feature maps. The symbol n_{ch1} denotes the number of feature map channels at the first level.

Three loss functions were explored to quantify the discrepancy between the model's predicted output and the actual target (ground truth). Specifically, metrics of mean absolute error (MAE) and mean squared error (MSE) were employed to measure the

L1 or L2 distance in either the spatial or frequency domain.

The loss function measuring the L1 distance in the image domain, $\mathcal{L}_{\text{Img-L1}}$, is defined as:

$$\mathcal{L}_{\text{Img-L1}} = \frac{1}{N} \sum_{i=1}^{N} |\mathbf{I}_{ns} - \hat{\mathbf{I}}_{ns}| \tag{4.1}$$

where $\mathbf{I}_{ns}$ and $\hat{\mathbf{I}}_{ns}$ represent the ground-truth and network-estimated final-position images, respectively; N is the total number of image pixels. The loss function measuring the L1 distance in the Fourier domain, $\mathcal{L}_{\text{F-L1}}$, is defined as:

$$\mathcal{L}_{\text{F-L1}} = \frac{1}{N} \sum_{i=1}^{N} |\mathcal{F}_2\left(\mathbf{I}_{ns}\right) - \mathcal{F}_2\left(\hat{\mathbf{I}}_{ns}\right)| \tag{4.2}$$

where $\mathcal{F}_2$ is the 2D Fourier transform operator. The loss function measuring the L2 distance, $\mathcal{L}_{\text{L2}}$, is defined as:

$$\mathcal{L}_{\text{L2}} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{I}_{ns} - \hat{\mathbf{I}}_{ns}\right)^2 \tag{4.3}$$

According to Parseval's theorem, and assuming all other training settings are constant, the L2 loss in both the image and Fourier domains should theoretically be equivalent.

For convenience, the UNet models trained with the three loss functions, $\mathcal{L}_{\text{Img-L1}}$, $\mathcal{L}_{\text{F-L1}}$, and $\mathcal{L}_{\text{L2}}$, are indicated as $\text{NN}_{\text{Img-L1}}$, $\text{NN}_{\text{F-L1}}$, and NN_{L2}, respectively.

4.1.2 Cartesian dataset

The main objective of this chapter is to validate the feasibility of deep learning-based intra-frame motion compensation techniques for reducing motion-related imaging errors in Cartesian cine-MRI. Therefore, the discussion and demonstration primarily focus on fully sampled *linear* Cartesian dataset as a case example, with only single-channel MRI included for simplicity.

In clinical practice, Cartesian MRI scanning can be accelerated by selectively skipping certain phase encoding lines in k-space to address the motion-related imaging errors, as scan time is approximately proportional to the number of time-consuming phase-encoding steps in k-space (see Chapter 2). Considerable efforts in undersampled MRI reconstruction have been directed toward mitigating aliasing artifacts [130], a major issue arising from violations of the Nyquist criterion [160] due to such omissions. To better reflect clinical realities in Cartesian cine MRI, this chapter further investigates the potential of the network's applicability for simultaneous undersampled MRI reconstruction and intra-frame motion compensation. Accordingly, a dataset for motion compensation in undersampled *linear* Cartesian MRI was generated.

The sub-Nyquist k-space sampling strategy was implemented following the specification outlined by Hyun, Chang Min, et al. [130]. First, uniform undersampling was applied in k-space along the phase-encoding axis, with a predefined acceleration factor, acc. The Poisson summation formula indicates that the T-periodic summation of a function f is expressed as discrete samples of its Fourier transform $\hat{f}$ with the sampling distance $1/T$:

$$\sum_{n=-\infty}^{\infty} f(x-nT) = \frac{1}{T} \sum_{n=-\infty}^{\infty} \widehat{f}\left(\frac{n}{T}\right) \mathrm{e}^{2\pi \mathrm{i}(n/T)x} \tag{4.4}$$

Consequently, for an $N \times N$ image matrix $\mathbf{I}(x, y)$, k-space subsampling by a factor of acc along the phase-encoding axis (i.e., y-axis), equivalent to a sampling interval of acc/N, produces the following fold-over image:

$$\mathbf{I}_{acc\text{-fold}}(x, y) = \sum_{j=0}^{acc-1} \mathbf{I}(x, y + \frac{jN}{acc}) \tag{4.5}$$

To address localization uncertainties caused by image folding, additional low-frequency lines were subsequently acquired. Fig. 4.3 illustrates reconstructed images obtained using different Cartesian sampling strategies. In Fig. 4.3 (a), a fully sampled coronal slice is presented, with the tumor located in the lower left lung. Fig. 4.3 (b) displays undersampled images with acceleration factors of $acc = 2$ and $acc = 4$, both with and without the inclusion of low frequency lines. Zero-padding is applied to the missing phase encoding lines. In the folded image produced by uniform undersampling, the tumor appears in both the left and right lungs, with the instance in the right lung being a folded artifact. As a result, the fold-over image corresponds to multiple plausible fully sampled images, with the tumor appearing on the left side, the right side, or both. Consequently, uniform undersampling can create uncertainty in identifying the true target location. This ambiguity introduces uncertainty in determining the true tumor location, which is intrinsically unresolvable by a neural network. The incorporation of a small number of low-frequency lines effectively circumvents this problem, as demonstrated in the far-right column of Fig. 4.3, where the reconstructed images clearly indicate the correct tumor position.

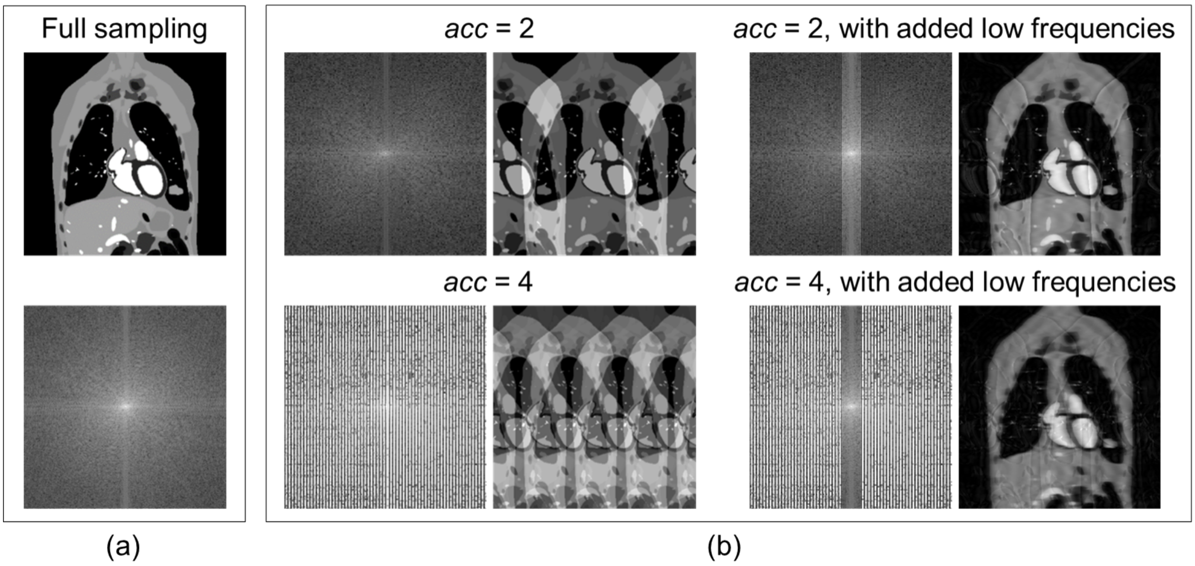

Figure 4.3: Cartesian sampling strategies. (a) Fully sampled image and corresponding k-space. (b) Undersampled k-space and images; the left columns show uniform undersampling with acceleration factors of $acc = 2$ and $acc = 4$, respectively, while the right columns show uniform undersampling with added low-frequency components.

The first 10 simulated patients were selected to create Cartesian datasets (see Section 3.4.1). For each patient, four original 2D+t cine-MR sequences were chosen from the 4D MRI digital anthropomorphic phantom: two sagittal and two coronal slices. One sagittal and one coronal slice containing the tumor centroid were selected. To enhance slice diversity, the other two slices were taken from non-tumor regions and specifically chosen to have distinct anatomical structures compared to the slices containing the tumor centroid. All frames were normalized by dividing them by their maximum magnitude values. The phase encoding was performed along the AP direction for sagittal slices and the left-right (LR) direction for coronal slices, both orthogonal to the main direction of intra-frame motion. The k-space matrix was filled from left to right with respect to the time steps.

For the fully sampled Cartesian dataset, the image matrices were generated as 512×512-pixel arrays, with a spatial resolution of 1 mm $\times$ 1 mm. The number of shots was set to 64 ($ns = 64$), i.e., the target was considered to remain stationary (or motion was negligible) while acquiring every 8 phase-encoding lines. To enable both the intra-frame motion compensation and denoising capabilities simultaneously, the input was the SNR $= 10$dB motion-corrupted image, used to predict the corresponding noiseless final-position image as output.

To more closely represent clinical conditions, for the undersampled Cartesian dataset, the image matrices were generated as 256×256-pixel arrays, with a spatial resolution of 1.5 mm $\times$ 1.5 mm. The acceleration factor was $acc = 4$, with 18 additional low-frequency lines acquired, as demonstrated by an example in the bottom right of Fig. 4.3. The number of shots was set to 82 ($ns = 82$), each shot corresponding to one single phase-encoding line. To enable intra-frame motion compensation, undersampled image reconstruction, and denoising simultaneously, the input-output pair was the SNR $= 10$dB motion-corrupted undersampled image and the corresponding noiseless final-position image.

A total of 14 intra-frame motion patterns were applied to simulate motion-corrupted frames based on each original cine-MR sequence. Consequently, the datasets included 11200 (10 patients $\times$ 4 slices $\times$ 20 frames $\times$ 14 patterns) input-output pairs, with data from eight randomly selected patients used for training (Patient 01, 03, 04, 05, 07, 08, 09) and validation (Patient 10), and the remaining 2 patients (Patient 02, 06) for testing. The images were represented as complex numbers and normalized by dividing them by the maximum magnitude value of the input before being fed into the network.

4.1.3 Evaluation Method

The effectiveness of the models was evaluated by comparing their outputs to the ground truth. Image quality enhancements were quantitatively assessed using MSE and MAE. Additionally, to evaluate target localization accuracy, the gross tumor volume (GTV) contours were generated for all sagittal frames containing tumors in the testing datasets, following the clinical MR-Linac procedure for online structure tracking.

In clinical practice, a preview cine MRI scan is acquired before treatment to select a tracking reference frame, denoted as $\mathbf{I}_{\text{ref}}$. During treatment, live cine MRI frames are aligned to $\mathbf{I}_{\text{ref}}$ using deformable image registration, and the GTV contour defined in the reference frame is propagated [30]. Similarly, in this work, $\mathbf{I}_{\text{ref}}$ and its corresponding GTV segmentation were defined: $\mathbf{I}_{\text{ref}}$ was directly selected from the original sequences, while its GTV was obtained by identifying the corresponding slice and frame from the 4D CT realistic tumor files and further processing it into a binary image. The DVF from $\mathbf{I}_{\text{ref}}$ to the floated frame was then computed utilizing the optical-flow algorithm [85]. Finally, the GTV was obtained by deforming the GTV of the reference frame based on the computed DVF.

The GTV contours were quantitatively compared using the Dice similarity coefficient (DSC) and the 95th percentile Hausdorff distance (HD$_{95}$) [161]. The DSC between two finite point sets, A and B, is defined as:

$$\text{DSC} = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{4.6}$$

where $|A \cap B|$ represents the number of elements in the intersection of sets A and B; $|A|$ and $|B|$ denote the total number of elements in sets A and B, respectively. DSC values close to 1 indicate a better overlap of the GTV contours. The HD_{95} is expressed as:

$$\text{HD}_{95}(A, B) = \max\{h_{95}(A, B), h_{95}(B, A)\} \tag{4.7}$$

where h_{95} represents the 95th percentile of the distances from all points in A to their nearest neighbor in B, and is defined as:

$$h_{95}(A, B) = \text{percentile}_{95}\left\{\min_{b \in B} \|a - b\| \middle| a \in A\right\} \tag{4.8}$$

Here, $\|a - b\|$ is the Euclidean distance between points a and b. A lower HD_{95} signifies closer alignment of the GTV contour to the ground truth.

4.1.4 Saliency map

The interpretability of deep neural networks [162] is particularly critical in high-stakes domains, such as healthcare, as discussed in this study. One approach to facilitate explanation in image processing is to identify pixels that are particularly influential, by calculating the gradient of the loss function w.r.t individual pixels x of the input image:

$$M_s(x) = \frac{\partial \mathcal{L}(x)}{\partial x} \tag{4.9}$$

The resulting saliency map, $M_s(x)$, assesses whether the model behaves as expected and can potentially provide insights into the underlying mechanisms.

Hence, to visualize which regions in the motion-corrupted image or k-space contribute most to the model's inference, saliency maps were generated in both the image and Fourier domains for the networks. Specifically, saliency maps in the image domain were computed using the SmoothGrad technique [163], which sharpens the saliency map through stochastic approximation:

$$\bar{M}_s(x) = \frac{1}{n}\sum_{1}^{n} M_s(x + \delta); \quad \delta \sim \mathcal{N}(0, \sigma^2)) \tag{4.10}$$

where n represents the number of samples; δ denotes noise randomly sampled from a standard Gaussian distribution and added to the input pixel x of the motion-corrupted image; and $\bar{M}_s(x)$ refers to the resulting average saliency map. To obtain saliency maps in the Fourier domain, input motion-corrupted image tensors were converted to the frequency domain and loaded onto the device (GPU) for gradient computation.

These tensors were then converted back to the image domain before being fed into the network.

4.1.5 Implementation details

The model was built with the PyTorch library [164], trained, and tested on an NVIDIA Quadro P5000 GPU with 16 GB of memory. A hyper-parameter search was conducted to determine the optimal initial learning rate for each model, sampling from the set $\{1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}\}$. The selected learning rates were 1×10^{-4} for $\text{NN}_{\text{Img-L1}}$, 1×10^{-3} for $\text{NN}_{\text{F-L1}}$, and 1×10^{-4} for NN_{L2}. The learning rate was reduced by a factor of 0.8 if no improvement was observed over 12 consecutive epochs. The Adam [165] optimizer was employed for all training processes, with a consistent batch size of 6 for all models.

4.2 Results

Unless otherwise specified, the results in this section are based on the fully sampled dataset.

To evaluate the network's inference speed, the average time required to estimate the final-position image was measured across the testing dataset, resulting in a measurement of 6.3 ms per frame.

Fig. 4.4 illustrates the training and validation losses for the UNet with three different loss functions, where the L2 loss is plotted on a logarithmic scale to highlight subtle differences. $\text{NN}_{\text{F-L1}}$ demonstrates a relatively larger discrepancy between the training and validation datasets compared to the other models. However, all validation loss curves eventually converge to a steady, horizontal line by the end of the training process. During the inference stage, the weights from Epoch 100 of all three models were loaded for testing.

To assess the models' performance from multiple perspectives, intra-frame motion was categorized into three scenarios: (i) *Static*, where the target remains stationary throughout the entire acquisition period; in this case, an ideal model should not introduce any positional changes to the target and should focus solely on image denoising; (ii) *Normal*, where the average intra-frame motion speed falls within the range of the published motion observations and remains below the maximum lung tumor speed reported to be 72.6 $\pm$ 22.5 mm/s [99] in the literature; (iii) *Extreme*, which is unlikely to occur in reality but was constructed to compel the network to prioritize the dynamic mechanism and prevent potential extrapolation regarding the motion amplitude.

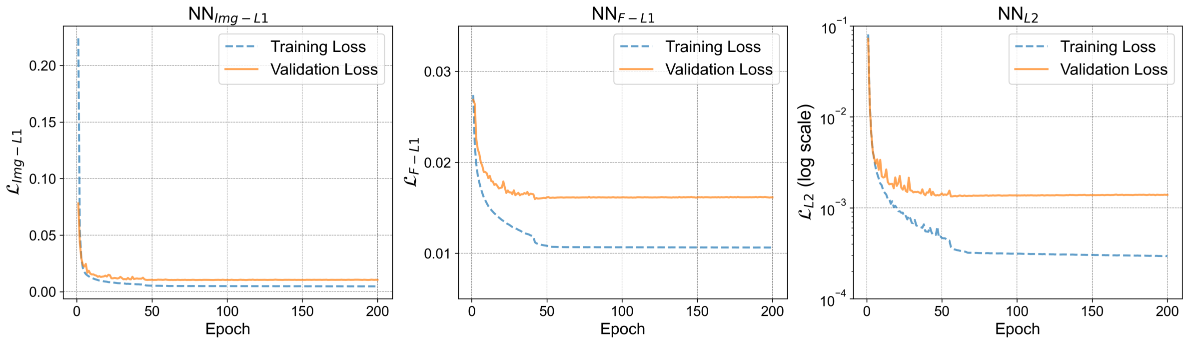

Figure 4.4: Training and validation loss curves for $NN_{Img\text{-}L1}$, $NN_{F\text{-}L1}$, and NN_{L2}. Note: $\mathcal{L}_{L2}$ values for NN_{L2} are plotted on logarithmic scale. This figure is adapted from material originally published in [137].

A comparison of the MSE and MAE values obtained from all testing frames is presented in Fig. 4.5, grouped by the three scenarios. Overall, the models significantly reduced imaging errors when compared to the ground truth across the testing dataset. $NN_{Img\text{-}L1}$ demonstrated a slight tendency towards a superior performance over the others in terms of both MAE and MSE. The results under the *Static* scenario highlight the models' denoising capabilities. For the *Normal* motion scenarios, applying $NN_{Img\text{-}L1}$, $NN_{F\text{-}L1}$, and NN_{L2} resulted in a decrease of median MSE (MAE) to 4.7% (10.5%), 6.2% (15.9%), and 12.0% (21.8%) of their initial values, respectively. In the *Extreme* scenario, a wider range of MAE or MSE variations was observed as anticipated. Nonetheless, all three models performed comparably well, with median MSE (MAE) values reduced to below 10% (18%) of their initial values, indicating the effective mitigation of intra-frame motion deterioration effects.

Fig. 4.6 and Fig. 4.7 present a comparison between representative motion-corrupted images and the network-estimated final-position images obtained from the testing dataset. The results indicate that applying UNets significantly improved image quality, with the network-estimated target positions demonstrating superior accuracy compared to those derived from the motion-corrupted images, particularly in the tumor, cardiac regions, and abdominal structures. The image noise was substantially mitigated. Among the models, $NN_{Img\text{-}L1}$ exhibited better image contrast restoration than $NN_{F\text{-}L1}$ and NN_{L2}, with pixel values more closely matching the ground truth in adipose and muscle tissues. Nevertheless, compared to imaging errors (target positioning errors and imaging blur), contrast inaccuracies were not considered critical in MRgRT and could be easily addressed by adjusting the intensity histograms.

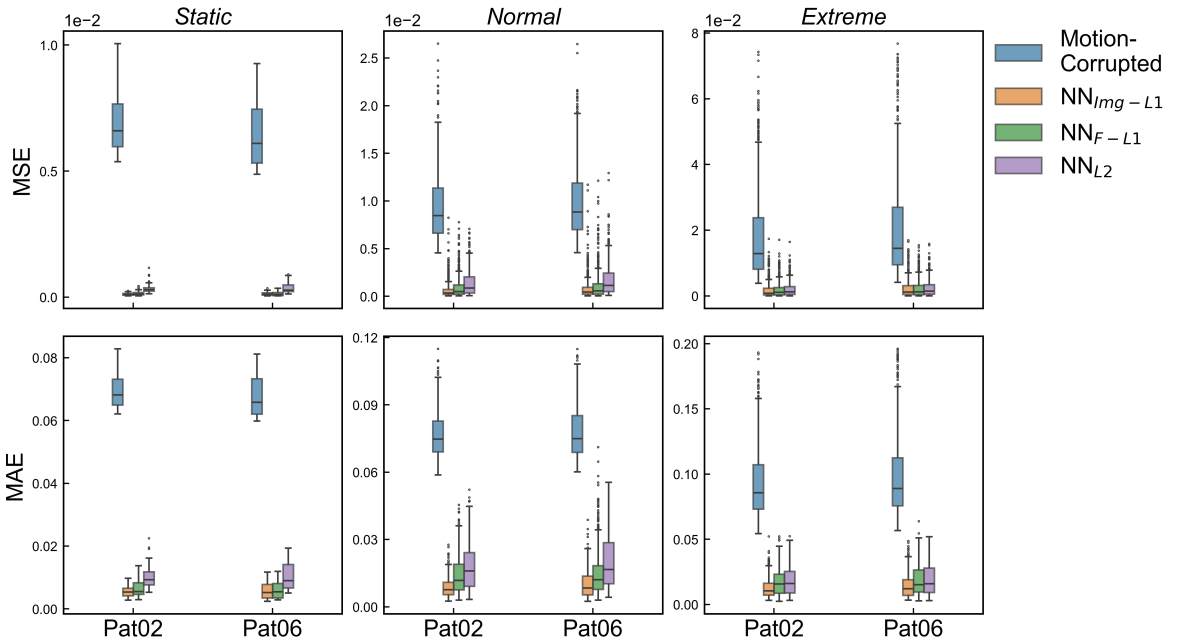

Figure 4.5: Box plots comparing the MSE (top) and MAE (bottom) of all testing frames before and after intra-frame motion compensation. Testing frames are categorized into three motion scenarios: *Static*, *Normal* and *Extreme*. Note: all images are normalized to the range [0, 1], and the y-axis scales vary across subfigures. This figure is adapted from material originally published in [137].

In particular, the tumor position in Fig. 4.6 was accurately corrected by the networks and was in close agreement with the ground truth. By comparing the motion-corrupted image to the reference final-position image in Fig. 4.6 and Fig. 4.7, it is evident that the cardiac regions experienced substantial intra-frame deformation during the frame acquisition. Nonetheless, all three compensation models were able to estimate the precise anatomical structure positions and shapes corresponding to the moment when the acquisition was completed. Additionally, the reduction in k-space discrepancies relative to the ground-truth also reflects a successful compensation of intra-frame motion by the models.

Target localization accuracy was evaluated with the slices in the testing dataset where the tumor centroid was located. The results were classified into three categories according to the GTV center of mass (COM) shift of the motion-corrupted image from the ground-truth: *Small*, for COM shift $\leq$ 2 mm; *Medium*, for 2 mm $<$ COM shift $\leq$ 5 mm; *Large*, for 5 mm $<$ COM shift $<$ 8 mm. Cases where the COM shift $>$ 8 mm were excluded, as in these cases, the intra-frame tumor motion speed exceeds the highest velocity observed in clinical studies, which is not realistic.

Fig. 4.8 and Table 4.1 present the evaluation results for all testing slices containing tumors, where the GTV contours of motion-corrupted and network-output images are compared with the ground truth using DSC and HD_{95}. The findings underscore the clear benefits of applying intra-frame motion compensation. All models achieved a significant improvement in DSC, with medians in each category exceeding 95%. The overall median DSC increased by 7 percentage points, from an initial value of 89%. Among the three models, $NN_{Img\text{-}L1}$ exhibited slightly better performance in the *Small* and *Medium* categories, while $NN_{F\text{-}L1}$ demonstrated a marginally higher median DSC in the *Large* category. Moreover, the networks reduced the median HD_{95} from 4.1 mm to 1.4 mm. These results indicate that, despite minor performance variations across different categories of intra-frame motion amplitude, all models are effective in compensating intra-frame motion, showcasing strong potential to eliminate target positioning errors within Cartesian cine-MRI for real-time motion management.

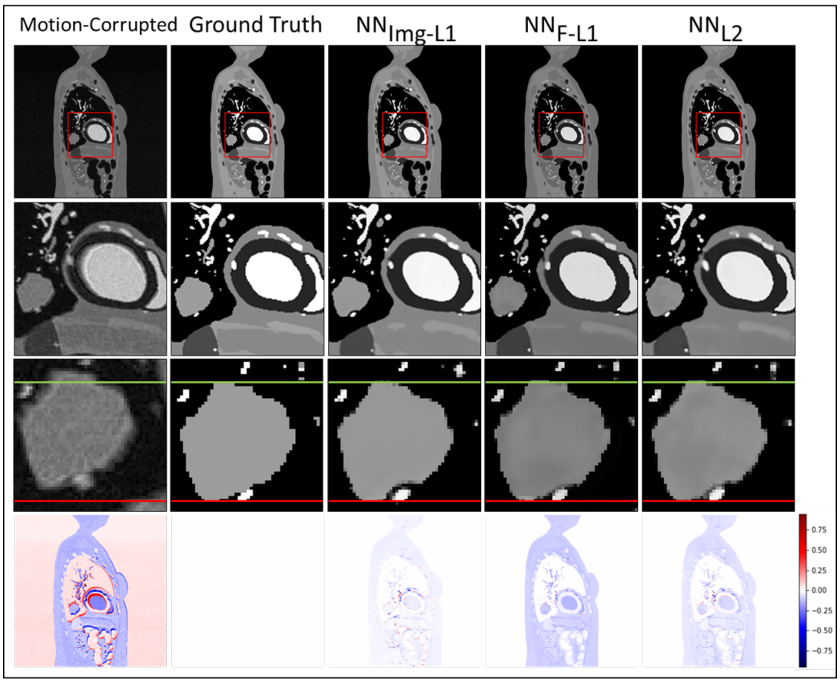

Figure 4.6: Comparison of representative sagittal frames before and after intra-frame motion compensation. From top to bottom: original image, zoomed-in cardiac region, magnified tumor area with reference lines marking the upper and lower boundaries of the ground-truth position, and image difference relative to the ground truth. This figure is adapted from material originally published in [137].

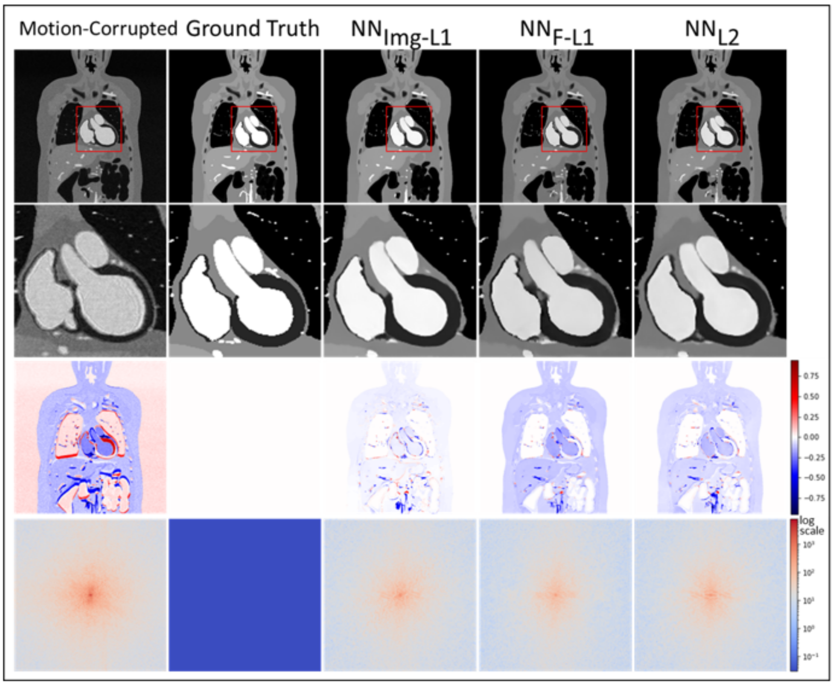

Figure 4.7: Comparison of representative coronal frames before and after intra-frame motion compensation. From top to bottom: original image, zoomed-in cardiac image, image difference, and the magnitude of k-space difference relative to the ground truth. The differences were computed by subtracting the ground truth. This figure is adapted from material originally published in [137].

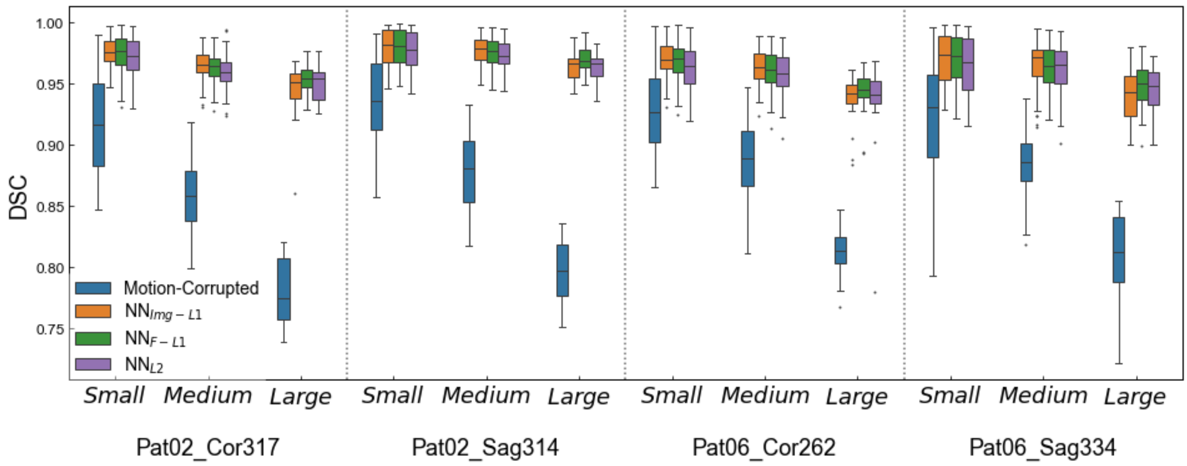

Figure 4.8: Box plot comparing target localization accuracies of the testing slices before and after intra-frame motion compensation. GTV contours of motion-corrupted and network-output images are quantitatively evaluated against the ground truth using DSC. Testing subjects are classified into three categories based on the GTV centroid shift: *Small*, *Medium*, and *Large*. This figure was originally published in [137].

Table 4.1: Quantitative evaluation of the measured GTV contours before and after intra-frame motion compensation. Median and [IQR] (interquartile range) of DSC and HD_{95} are reported for all testing slices containing tumors.

	DSC (%)				HD_{95} (mm)			
	Motion-Corrupted	$NN_{Img\text{-}L1}$	$NN_{F\text{-}L1}$	NN_{L2}	Motion-Corrupted	$NN_{Img\text{-}L1}$	$NN_{F\text{-}L1}$	NN_{L2}
Small	92.7 [5.4]	97.5 [2.2]	97.5 [2.5]	97.2 [2.6]	2.8 [2.1]	1.0 [0.4]	1.0 [0.4]	1.0 [1.0]
Medium	88.0 [4.4]	96.9 [2.1]	96.6 [2.3]	96.5 [2.2]	4.5 [2.3]	1.4 [1.0]	1.4 [1.0]	1.4 [1.2]
Large	80.8 [4.4]	95.0 [2.3]	95.5 [2.4]	95.2 [2.7]	7.1 [2.0]	2.0 [1.4]	2.0 [1.4]	2.0 [1.4]
Total	89.4 [8.1]	96.9 [2.6]	96.8 [2.6]	96.6 [2.7]	4.1 [3.3]	1.4 [1.0]	1.4 [1.0]	1.4 [1.2]

By altering the order of the frames in the original cine-MR sequences and utilizing the intra-frame motion pattern perturbation scheme proposed in Section 3.4.2.2, it is feasible to customize arbitrary synthetic yet realistic breathing motion curves, including dedicated intra-frame motion trajectories. Using this approach, GTV centroid motion curves were constructed for sagittal slices of Patient 02 and Patient 06 from the testing dataset, as shown by the red line in Fig. 4.9, which serves as the ground truth. The absolute GTV centroid positions derived from motion-corrupted images and the network-estimated final-position images are compared to the ground truth. As illustrated in the figure, motion-corrupted results show that most frames exhibited an imaging latency of approximately 50% of the frame acquisition time. However, a longer time delay was evident for certain frames of Patient 06, particularly Frames 10, 11, and 13. This can be attributed to the potential degradation of image quality caused by motion artifacts and noise, which adversely affects the accuracy of the optical flow algorithm. The three network-estimated results overlap well with the ground truth across all cases, effectively correcting GTV position offsets. The only exception occurred in Frame 13 for Patient 06, where the optical flow algorithm failed to precisely contour the tumor in the $NN_{Img\text{-}L1}$ and $NN_{F\text{-}L1}$ estimated images. The target positioning errors were negligible or completely absent in cases with a very shallow breathing mode, such as in Frame 13 to 18 for Patient 02.

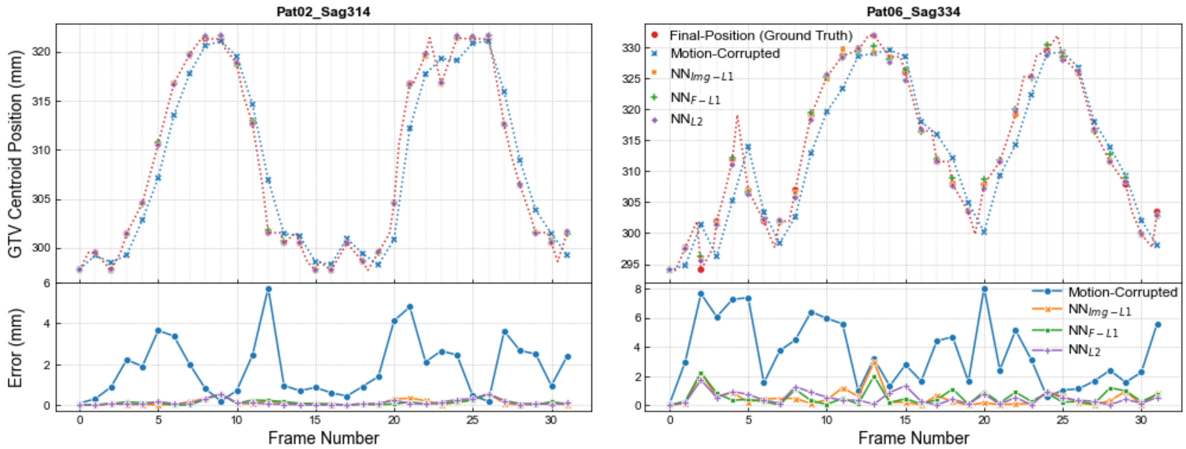

Figure 4.9: GTV centroid position comparison curve. The constructed breathing motion curves, including intra-frame motion trajectories, are depicted by the red line, while the red dots indicate the ground-truth GTV centroid position at the moment the frame acquisition is terminated. Results before and after motion compensation are displayed: motion-corrupted results are shown in blue, $NN_{Img\text{-}L1}$ in yellow, $NN_{F\text{-}L1}$ in green, and NN_{L2} in purple. The difference curves relative to the ground truth are presented in the lower panels. This figure was originally published in [137].

To identify the regions in the input motion-corrupted image or k-space that exert the greatest influence on the model's inference, saliency maps of the loss function with respect to the input were generated in both the image and Fourier domains for the three models. The overlaid saliency maps of representative testing patients are shown in Fig. 4.10. On the one hand, the right part of k-space corresponding to the later-acquired data is highlighted in the heat map, representing a large contribution to the final results; on the other hand, saliency maps in the image domain indicate a primary focus on the edges of the moving structures. In particular, the models are capable of detecting the edges at their final positions during the frame acquisition, which are imperceptible to humans, as evidenced by the coronal slice, where the model-highlighted liver edge deviates from the edge perceived by visual observation.

Fig. 4.11 illustrates representative examples of imaging error reduction in undersampled Cartesian cine-MRI, where the network was tasked with simultaneously performing intra-frame motion compensation, undersampled image restoration, and image denoising. NN_{L2} was selected as the correction model, and the corresponding input motion-corrupted undersampled images are compared with the images processed by the network. The results demonstrated significant advantages of implementing NN_{L2}: aliasing artifacts caused by sub-Nyquist k-space sampling were effectively suppressed; structural localization was accurately corrected, as observed in tumor regions and other anatomies affected by respiratory motion; and image noise reduction was appreciable.

Consistent with the observations in Fig. 4.6 and Fig. 4.7, NN_{L2} generated images exhibited minor contrast inconsistencies with the ground-truth. While this discrepancy was far less critical for MRgRT than imaging errors, it could be effectively resolved through histogram matching techniques. The difference images demonstrated values that converge more closely to zero after this correction.

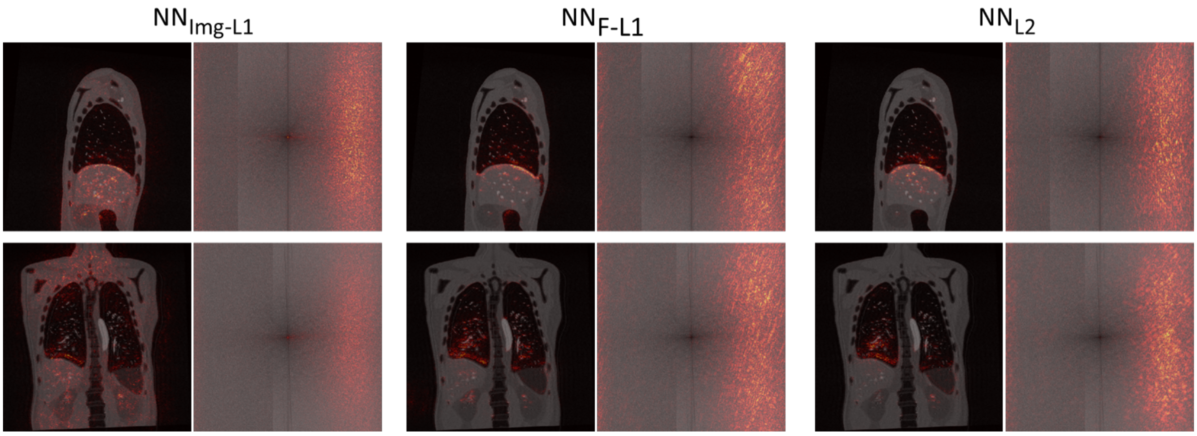

Figure 4.10: Overlaid saliency map in the image (left) and Fourier (right) domain for model $NN_{Img\text{-}L1}$ (top), $NN_{F\text{-}L1}$ (middle) and NN_{L2} (bottom). This figure is adapted from material originally published in [137].

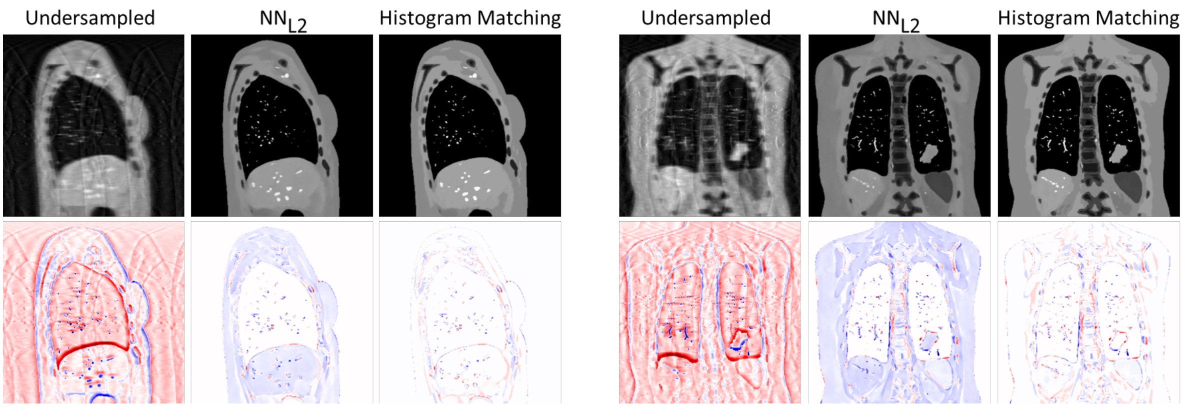

Figure 4.11: Imaging error reduction in undersampled Cartesian cine-MRI. Motion-corrupted undersampled images (left) are compared with NN_{L2} processed images, both with (right) and without (middle) histogram matching. Difference images (bottom) were computed by subtracting the ground truth.

4.3 Discussion

This chapter investigates the feasibility of reducing imaging errors in Cartesian cine-MRI by implementing deep learning-based intra-frame motion compensation techniques.

The motion-corrupted image decomposition experiment depicted in Fig. 4.1 reveals that, despite being obscured by dominant LFC information, the contours of the structures' ground-truth real-time positions are encoded within the motion-corrupted images, corresponding to the later acquired HFCs in the Fourier domain. This finding suggested the selection of a convolutional neural network for the task, given its exceptional capability in extracting frequency-domain information. Considering the application scenario involving *linear* phase encoding Cartesian k-space sampling trajectories, a suitable compensation model must preserve the later-acquired HFCs while processing the LFC-associated patterns. The UNet architecture, with its skip connections that enable the reuse of fine-grained deep features, stands out as a particularly promising model for this purpose.

To this end, UNet models were trained using the generated Cartesian datasets to estimate the final-position image directly from the motion-corrupted inputs. The models provided simultaneous intra-frame motion compensation, image denoising, and mitigation of aliasing artifacts for undersampled images. Three types of loss functions were investigated for performance comparison.

The inference time plays a vital role in enabling the practical implementation of this technique for real-time motion management. The network required approximately 6.3 ms to complete the motion compensation for a 512×512 image, which was clinically acceptable, as it was significantly shorter than current clinical Cartesian cine-MR frame acquisition time, such as 4 Hz (i.e., 250 ms/frame) in the ViewRay MRIdian system [30]. Furthermore, this speed is highly dependent on the hardware configuration and the matrix size of the input: with ongoing advancements of GPU computing power, the actual processing time is expected to be further reduced.

The models were comprehensively evaluated on the testing dataset, demonstrating their ability to significantly reduce imaging errors. This was reflected in improved image quality metrics such as MSE or MAE, as well as enhanced GTV contour measures, including DSC and HD_{95}. Specifically, for the testing dataset analyzed in GTV contouring, the median DSC increased from 89% to 97%, while the HD_{95} dropped from 4.1 mm to 1.4 mm. Additionally, in Fig. 4.6 and Fig. 4.7, substantial deformations in the cardiac region were observed within a single cine-MR frame acquisition. Nonetheless, the models exhibited the capability to accurately estimate the anatomical structure at the moment the acquisition was completed, highlighting their potential advantages for real-time MR imaging of cardiac function.

The three models exhibited slight performance variations across different motion amplitude categories: $\mathrm{NN}_{\text{Img-L1}}$ excelled in the *Small* and *Medium* cases, whereas $\mathrm{NN}_{\text{F-L1}}$

demonstrated a bit higher median DSC values in the *Large* category. This outcome aligns with expectations, as the input-output image pairs of the network are normalized, limiting the absolute prediction errors to less than 1 per pixel. Consequently, the L2 loss is less sensitive to outliers compared to the L1 loss.

In Fig. 4.9, the GTV centroid position derived from the motion-corrupted image generally corresponds to the position at half of the frame acquisition time. This is consistent with the findings of Borman et al. [88] and Riederer et al. [143], which demonstrate that the target position is primarily determined by the moment when the central k-space profile is acquired. As a result, a linearly and fully acquired Cartesian readout k-space trajectory leads to an imaging latency of approximately 50% of the acquisition time. Notably, the network-estimated positions overlap well with the ground truth, showing a clear benefit.

The saliency maps of the motion-corrupted input in Fig. 4.10 highlight the far right region of k-space as well as the edges of the moving anatomical structures, with these detected edges representing their final positions, which may differ from those observed visually. This makes it more transparent that the models have learned to identify and extract information from the later-acquired frequency components, which in turn guides the alignment of the corresponding image features acquired earlier. This behavior is noteworthy and particularly important for addressing concerns regarding the potential and reliability of deep learning approaches for clinical implementation.

In addition to reducing motion-related imaging errors, the network is highly versatile, demonstrating the ability to perform multiple tasks simultaneously. This is exemplified by the undersampled Cartesian MRI experiment (see Fig. 4.11), where the model effectively carried out intra-frame motion compensation, suppressed aliasing artifacts, and denoised the image. Depending on clinical needs, other functionalities can be incorporated, such as training the model to directly output segmentation results for GTVs or OARs without localization errors.

4.4 Conclusions

This chapter explores the potential of deep learning-based intra-frame motion compensation techniques to reduce imaging errors in Cartesian cine-MRI. UNets with three types of loss functions were successfully trained to estimate the exact noiseless final-position image from the motion-corrupted input. The models led to an evident image quality and GTV position accuracy enhancement, confirmed by a decreased image MSE/MAE and an improvement in terms of GTV DSC and HD_{95}. Saliency maps indicated that the models learned to utilize later-acquired frequency components to improve the convergence of the earlier-acquired corresponding image features. The networks' versatility was further demonstrated in the undersampled Cartesian MRI experiment, where the aliasing artifacts were effectively mitigated. These findings highlight the promising capability of deep learning-based intra-frame motion compensation techniques to improve imaging accuracy in Cartesian cine-MRI, paving the way for their application in real-time motion management.

Chapter 5

INTRA-FRAME MOTION COMPENSATION FOR RADIAL CINE-MRI

5.1 Method and materials

5.1.1 Overall workflow

As schematically depicted in Fig. 5.1, radial MR sequences enhance the frame rate by reducing the stride of the sliding reconstruction window. However, the imaging latency manifested by target positioning errors is independent of the frame rate and instead correlates with the temporal coverage of spokes within the reconstruction window. Due to single-frame acquisition and the physiological motion occurring on similar time scales, the acquired k-space data within the window may comprise signals from the target at varying positions. For instance, window $M+3$ in Fig. 5.1 consists of ns radial spokes corresponding to time steps from t_1 to t_{ns}. Throughout the acquisition period, the target transitions from positions p_1 to p_{ns}. By utilizing a tailored set of sampling matrices specific to the online radial k-space readout trajectory, this motion-dependent data acquisition process can be simulated with the procedure outlined in Section 3.1, where corresponding complex-valued radial spokes in the frequency domain are sequentially incorporated into the k-space arrays constructed over the time steps.

Conventionally, the acquired samples within the reconstruction window are directly reconstructed into an image with 2D inverse NuFFT, resulting in imaging errors, as depicted by the motion-corrupted image in Fig. 5.1. Unlike existing work on highly undersampled image reconstruction that addresses this issue by reducing the window width, it is hypothesized that the spokes sampled earlier, regardless of their temporal distance from the last time step t_{ns}, still contribute to the precise recovery of the image. In this study, without compromising the window width, the intra-frame motion compensation model TransSin-UNet attends over all spokes as well as their associated spatial and temporal information in the k-space, and derives the final-position image at the time of the last shot.

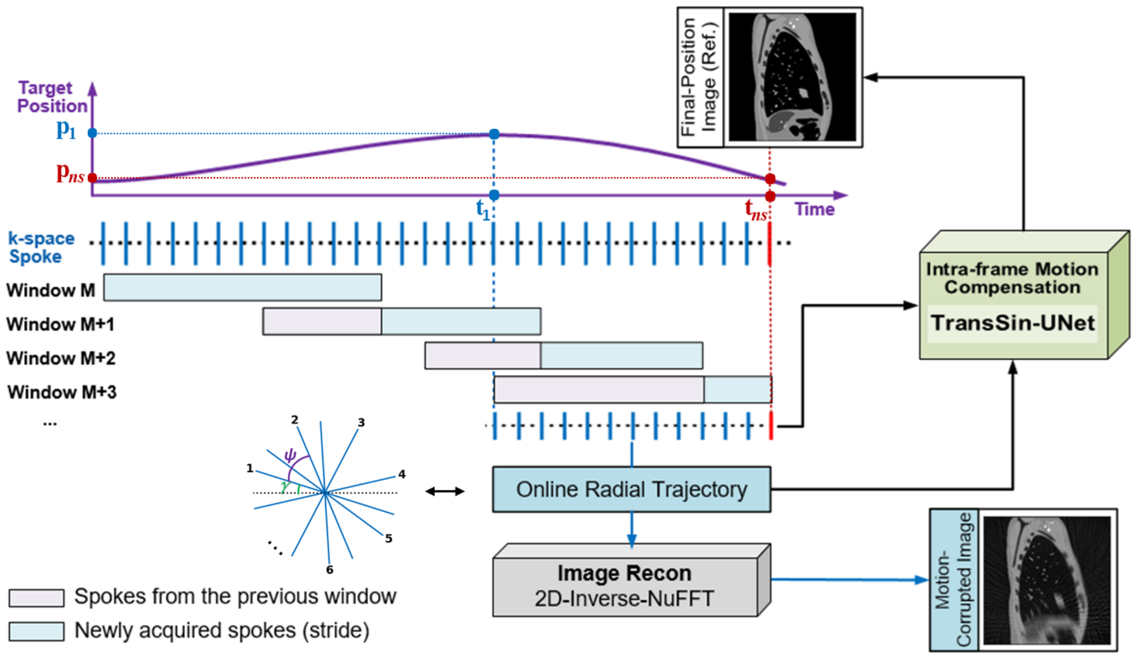

Figure 5.1: Schematic diagram of the motion-dependent radial sampling and the overall framework of the proposed method. This figure is adapted from material originally published in [166].

5.1.2 Intra-frame motion compensation network: TransSin-UNet

5.1.2.1 TransSin-UNet model

Convolutional neural networks excel at identifying information associated with specific frequency ranges, making them particularly effective for Cartesian problems. Unlike Cartesian cine-MRI, where later acquired data are concentrated in specific high- or low-frequency regions (along the phase encoding direction) of the k-space, such as HFC in *linear* sampling and LFC in *high-low* sampling, radial sampling presents fundamentally different complexities.

Firstly, each spoke in the radial trajectory passes through the origin and uniformly spans both high and low frequencies in the Fourier domain. Secondly, the sliding window approach introduces variability, as the first spoke within a reconstruction window can originate at any arbitrary position, defined by its starting angle γ, resulting in unique trajectory coordinates for each frame. Furthermore, in contrast to *linear* radial trajectories, where temporally close spokes are also spatially close, *(tiny) golden angle* acquisitions interleave newly acquired spokes with previously acquired ones. Consequently, the talks among the spokes must be modeled with consideration of both spatial and temporal adjacency. However, CNNs typically leverage spatial locality by restricting neuron connections to neighboring regions, resulting in a limited receptive field that is inadequate for attending long-distance interactions. To address the complexities of the radial problem, alternative architectures are required. Attention mechanisms, which can be viewed intuitively as a sophisticated form of CNN with adaptive and learnable receptive fields, have emerged as a promising solution.

Therefore, in this work, TransSin-UNet is proposed as an intra-frame motion compensation model especially tailored to reduce motion-related imaging errors in radial cine-MRI. As shown in Fig. 5.2, the model integrates a sinogram transformer encoder (referred to as SinTE) and a UNet to perform dual-domain operations. On the one hand, imaging errors caused by intra-frame motion of the target originate in the Fourier domain, therefore, an intuitive strategy to mitigate these errors involves processing the acquired data directly in the k-space, aligning the temporal spokes with those of the ground-truth image. This sequence-to-sequence regression is facilitated by the transformer encoder, the prominent architecture of choice in establishing long-range dependencies among the input, leveraging its self-attention mechanism. On the other hand, given that the downstream tasks of MRgRT rely on cine-MR image data, the UNet refines the reconstruction through a pixel-level fine-tuning within the image domain, facilitated by its exceptional capacity to capture intricate local details.

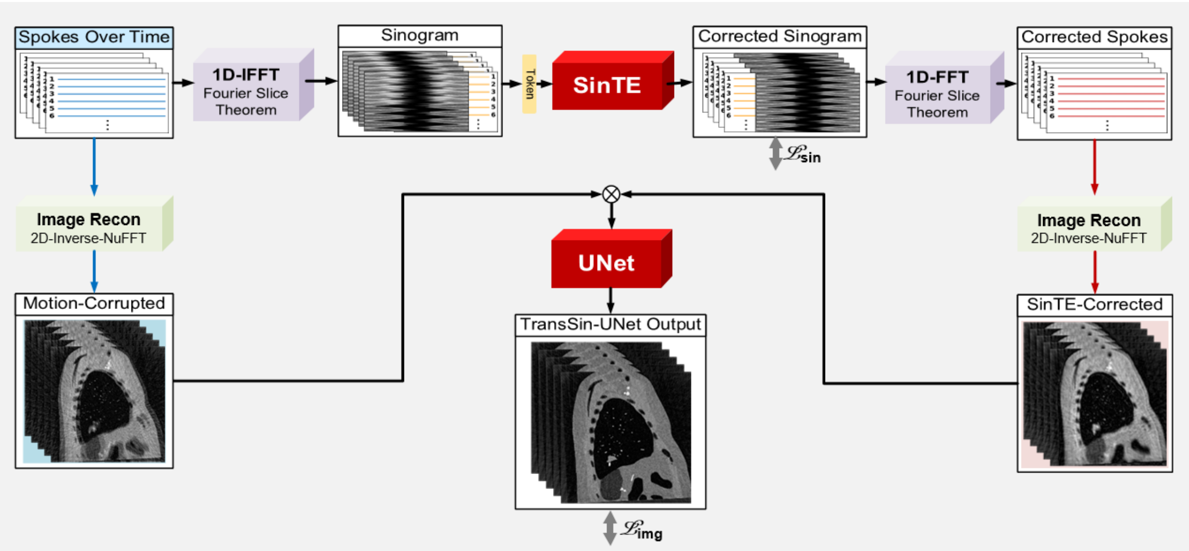

Figure 5.2: TransSin-UNet model. The architecture integrates a sinogram transformer encoder (SinTE) with a UNet to perform dual-domain processing. SinTE learns spatial-temporal dependencies among sinogram representations of radial spokes, performing sequence-to-sequence regression in the projection domain to align the temporal spokes with the ground-truth; The UNet performs pixel-level fine-tuning within the image domain. This figure is adapted from material originally published in [166].

As illustrated in Fig. 5.2, the complex-valued radial spokes are first reorganized sequentially based on their acquisition time steps. Considering the power spectrum characteristics of medical images, where the central k-space exhibits significantly higher energy than the peripheral regions, the values along each spoke span a wide range of magnitudes. Directly using these values as input may lead to poorly conditioned gradients of the non-linear activation functions in the transformer encoder, potentially hindering convergence. To address this, a mapping of the spoke data from the frequency domain to the projection domain is considered, based on the Fourier projection-slice theorem.

The theorem states that a slice of the 2D Fourier transform of a function, taken along a line passing through the origin, is equivalent to the Fourier transform of the projection of the 2D function onto a parallel line. Therefore, it follows that:

$$\mathcal{F}_1^{-1}\mathbf{S}(\omega\cos\theta, \omega\sin\theta) = \int_{-\infty}^{\infty} \mathbf{S}(\omega\cos\theta, \omega\sin\theta)e^{2\pi i p\omega}d\omega = \mathcal{R}_\theta[f](p) \tag{5.1}$$

where $\mathcal{F}_1^{-1}$ represents the inverse 1D Fourier transform operator; $\mathbf{S}(\omega\cos\theta, \omega\sin\theta)$ signifies the radial k-space spoke at angle θ; and $\mathcal{R}_\theta[f](p)$ denotes the Radon transform, which computes line integrals and projects the image onto the line at angle θ.

Fig. 5.3 illustrates the conversion relationships between the spatial, projection and frequency domains as described by the Fourier projection-slice theorem. In the spatial domain, the projection axis (p-axis) forms an angle θ with the x-axis. The projection of the image onto the p-axis is computed by applying the Radon transform along a set of parallel lines perpendicular to the p-axis (blue lines). The result of this process is visualized in the projection domain as a graph of $\mathcal{R}_\theta[f](p)$, providing the line integral values as a function of position along p-axis. $\mathcal{R}_\theta[f](p)$ further corresponds to a line in the sinogram space, which compiles projections over a range of angles. A radial spoke at angle θ in k-space, represented along the ω-axis (parallel to the p-axis), can be viewed as a slice through the frequency domain. Since $\mathcal{R}_\theta[f](p)$ and $S(\omega)$ are 1-dimensional Fourier transform pairs, the frequency components can be mapped back to the projection domain, converting the spoke signal values to a scale range comparable to the original image intensity values.

Consequently, each spoke is inversely Fourier transformed to yield a representation in the projection domain of the image along its angle, known as its sinogram representation. This process reduces the dominance of central k-space values, ensuring a more balanced magnitude distribution across all input dimensions. With np representing the number of readout points sampled along each spoke, the real and imaginary parts of the sinogram representation for each spoke are stacked into a $2np$-dimensional vector, which is treated as a token of the input sequence. The token vectors then pass through the sinogram transformer encoder, which models their spatial-temporal correlations and generates the corrected sinogram as output. Afterward, each row of the output sinogram is converted to complex form and translated back to k-space using a 1D Fourier transform. Subsequently, the process involves simultaneous and parallel image reconstruction with (i) the original k-space spokes to obtain the motion-corrupted image, and (ii) transformer-encoder corrected spokes to obtain the SinTE-corrected image. The reconstruction is realized by 2D inverse NuFFT based on the individual k-space trajectory of each frame. Finally, with the real and imaginary parts represented as separate channels, the two complex-valued images are concatenated and fed into the UNet to estimate the real-time final-position image.

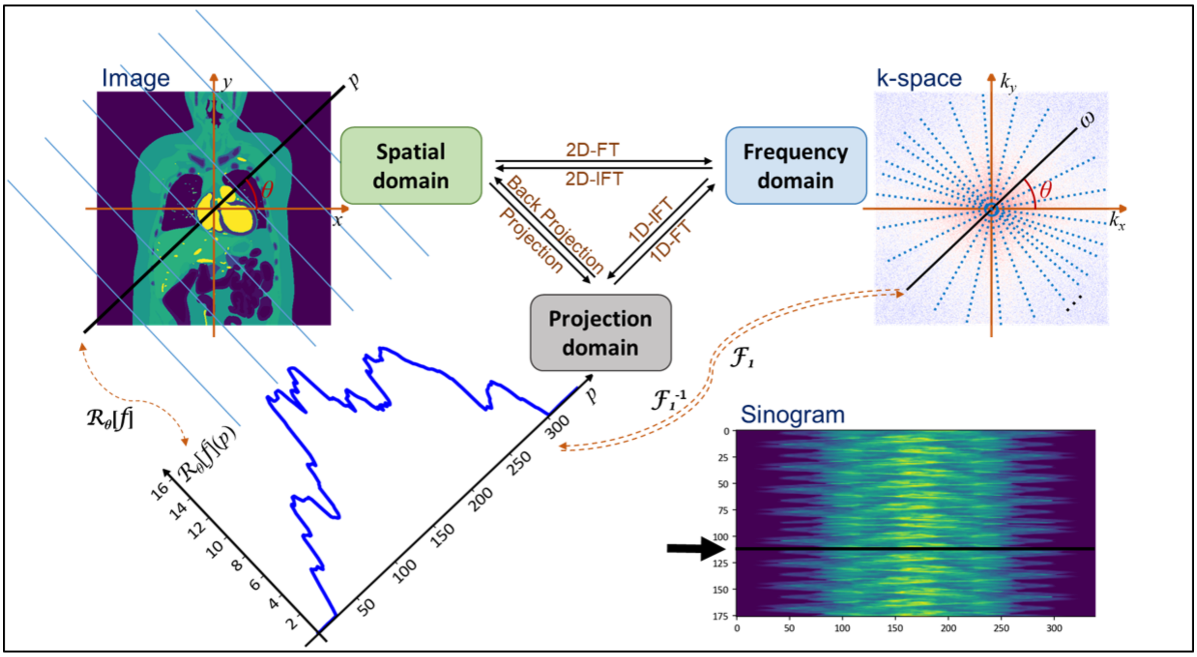

Figure 5.3: The relationship between the spatial, projection and frequency domains as described by the Fourier projection-slice theorem. $\mathcal{R}_\theta[f]$, $\mathcal{F}_1$ and $\mathcal{F}_1^{-1}$ represent the Radon transform, 1D Fourier transform (FT), and 1D inverse Fourier transform (IFT) operators, respectively. This figure was originally published in [166].

5.1.2.2 Joint loss function

To guide a stable training process of the network, a joint loss function $\mathscr{L}$ is defined as a weighed linear combination of the sinogram loss $\mathscr{L}_{\text{sin}}$ and the reconstructed image loss $\mathscr{L}_{\text{img}}$:

$$\mathscr{L} = \alpha \times \mathscr{L}_{\text{sin}} + \mathscr{L}_{\text{img}} \tag{5.2}$$

where α is the weight parameter. In this work, the L1 loss is used to quantify the discrepancy between the network output and the ground truth. Consequently,

$$\begin{aligned}\mathscr{L}_{\text{sin}} &= \sum_{\theta,p} \left|\mathbf{T}\left[\mathcal{R}f_{\text{motion}}\left(\theta,p\right)\right] - \mathcal{R}f_{\text{ref}}\left(\theta,p\right)\right|; \\ \mathscr{L}_{\text{img}} &= \sum_{x,y} \left|f_{\text{out}}\left(x,y\right) - f_{\text{ref}}\left(x,y\right)\right|\end{aligned} \tag{5.3}$$

where $\mathbf{T}$ represents the sinogram transformer encoder; $\mathcal{R}f_{\text{motion}}(\theta,p)$ signifies the motion-corrupted sinogram, i.e. the input of $\mathbf{T}$; $f_{\text{out}}(x,y)$ is the output image of the TransSin-UNet/UNet; and $\mathcal{R}f_{\text{ref}}(\theta,p)$ denotes the reference sinogram calculated from the ground-truth final-position image $f_{\text{ref}}\left(x,y\right)$.

5.1.2.3 Subnetwork: Sinogram transformer encoder (SinTE)

The architecture of SinTE is outlined in Fig. 5.4, consisting of the positional encoding, $N = 8$ identical transformer encoder blocks, and a linear output layer.

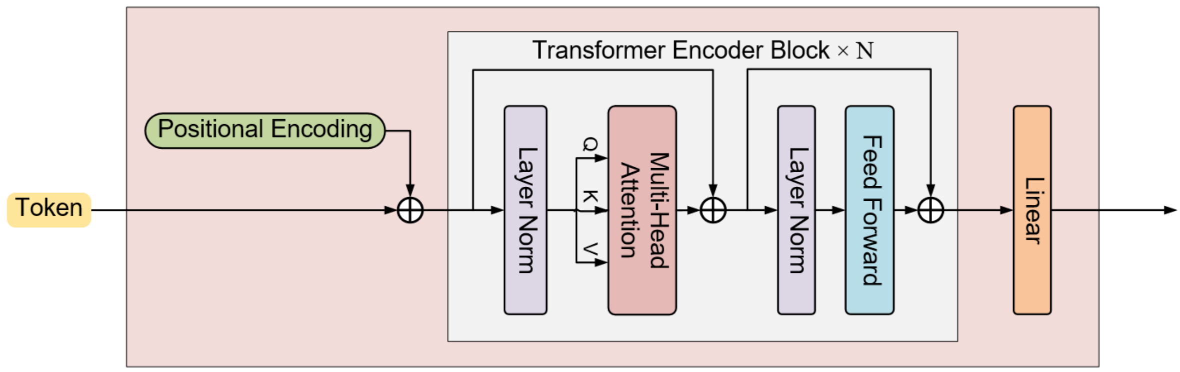

Figure 5.4: The architecture of the Sinogram Transformer Encoder. This figure is adapted from material originally published in [166].

Theoretically, positional encoding should be implemented to provide the transformer encoder with the position information of the spokes in both the spatial and temporal dimensions, which is closely tied to their dependencies. The relative or absolute temporal positions of the spokes within each frame are intuitively encoded based on their acquisition time step. However, in the spatial dimension, the absolute positions of the spokes are influenced by the random starting angle γ of the first shot for each frame. To simplify the encoding process, the spatial and temporal positional encodings are unified by considering only the relative spatial positions of the spokes, which depend exclusively on their acquisition time step under the condition of a constant angular increment of ψ between temporally consecutive spokes. To this end, based on the properties of sinusoidal functions, the position encoding matrix PE is defined as [167]:

$$\begin{aligned} PE\left(idx, 2i\right) &= sin\left(idx/10000^{2i/d_{model}}\right) \\ PE\left(idx, 2i+1\right) &= cos\left(idx/10000^{2i/d_{model}}\right) \end{aligned} \tag{5.4}$$

where $idx \in [1, 2, ..., ns]$ is the time step index of the token; $2i$ and $2i+1$ denote the dimensions; d_{model} is the total dimensionality of each spoke vector, which is set to $d_{model} = 2np$ in this work. The generated positional encoding matrix is visualized in Fig. 5.5. The function (Eq. 5.4) is hypothesized to enable the model to easily attend to the relative positions of the spokes. As for a fixed offset s, $PE(idx+s)$ can be represented as a linear transformation of $PE(idx)$. The obtained positional encoding values are directly added to their corresponding input tokens.

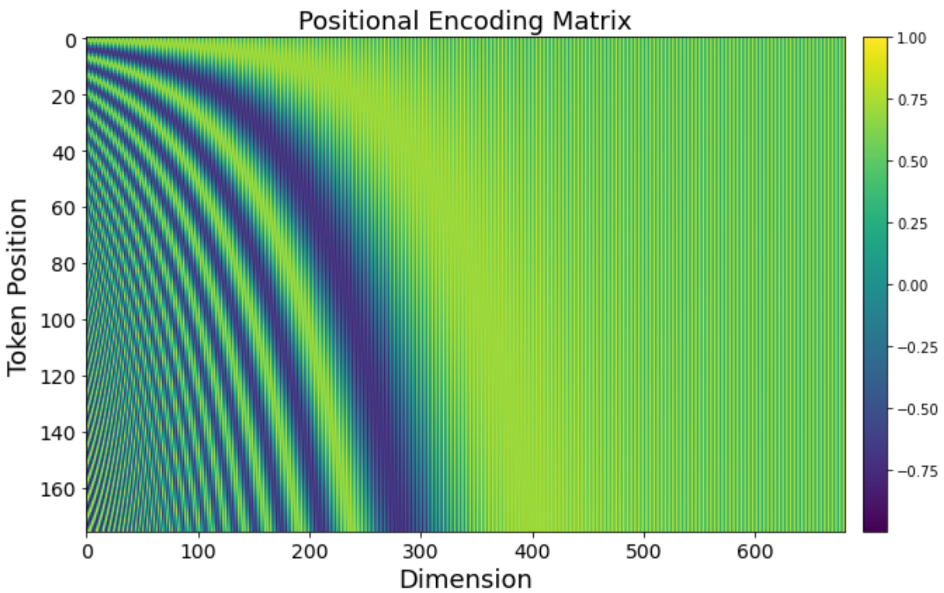

Figure 5.5: Visualization of the positional encoding matrix generated with sinusoidal functions.

Each identical block in the encoder comprises two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feedforward network (FFN). A residual connection is employed around each of the two sub-layers. To ensure stable gradient behavior during initialization and avoid issues such as exploding or vanishing, the pre-LN structure [168] is adopted, placing the layer normalization inside the residual connection.

In the self-attention mechanism, the input tokens are related through attention scores to compute a contextualized representation of the sequence. The attention operation for each head is defined as [167]:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{5.5}$$

where the query matrix $\mathbf{Q} = \mathbf{X}\mathbf{W}_q$, key matrix $\mathbf{K} = \mathbf{X}\mathbf{W}_k$, and value matrix $\mathbf{V} = \mathbf{X}\mathbf{W}_v$ are linear projections of the input $\mathbf{X}$ (formed by all input tokens), with $\mathbf{X}$, $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V} \in \mathbb{R}^{ns \times d_{model}}$. To allow the model to jointly attend to information from different representation subspaces across different positions, multi-head attention was employed:

$$\begin{gathered} \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, ..., \text{head}_h)\mathbf{W}^O \\ \text{where head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \end{gathered} \tag{5.6}$$

where $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$,$\mathbf{W}_i^V \in \mathbb{R}^{d_{model} \times d_k}$, and $\mathbf{W}^O \in \mathbb{R}^{hd_k \times d_{model}}$ are parameter matrices for the projection; h represents the number of attention heads, which is set to $h = 8$ in this work; $d_k = d_{model}/h$.

FFN consists of two linear transformations with a non-linear activation function in between. The inner-layer has the dimensionality of $d_{ff} = 1024$.

5.1.2.4 Subnetwork: UNet

The UNet subnetwork adopts a 4-level architecture. Each level, linking the contracting and expansive paths via skip-connections, comprises a double convolution block with 3×3 convolution kernels, followed by batch normalization and LeakyReLU activation. As previously detailed, the UNet processes concatenated inputs, consisting of the motion-corrupted image and the image reconstructed from the SinTE-corrected spokes, to generate the estimated final-position image. The real and imaginary parts of all complex-valued images are separated into two distinct channels, resulting in input and output channel numbers of 4 and 2, respectively.

The first level contains 32 feature channels, which are doubled sequentially at each subsequent level. Down-sampling within the contracting path is performed using 2×2 max pooling with a stride of 2, while up-sampling in the expansive path employs up-convolution [169] followed by concatenation. A 1×1 convolution operation layer serves as the linear output layer of the network.

5.1.2.5 Decomposition of online radial trajectories

The spatial-temporal information of the acquired data is recorded by the online radial k-space sampling trajectory, where consecutive spokes are arranged with a successive angular increment ψ. In this Chapter, ψ is set to either the *golden angle* ψ_{gold} or the *tiny golden angle* ψ_N.

The use of a sliding window, where the first shot within a given reconstruction window may begin at any arbitrary position (represented by γ), results in each frame possessing a distinct sampling trajectory. As depicted in Fig. 5.6, instead of storing the online trajectory coordinates for each frame individually, a unified default trajectory—starting at $0°$ and determined solely by ψ—is used in conjunction with the frame-specific random starting angle γ. The 2D inverse NuFFT employed in the model requires density correction factors specific to the trajectory to ensure uniform k-space sampling density. However, the online calculation of DCFs can be computationally expensive. With the default trajectory, the corresponding default DCFs can be precomputed. Consequently, the samples are populated onto the default k-space trajectory, and the image is reconstructed, which is equivalent to a counterclockwise rotation in the frequency domain. According to the rotational invariance property of the Fourier transform (Appendix A.2), the true image can then be recovered by simply rotating γ clockwise in the spatial domain.

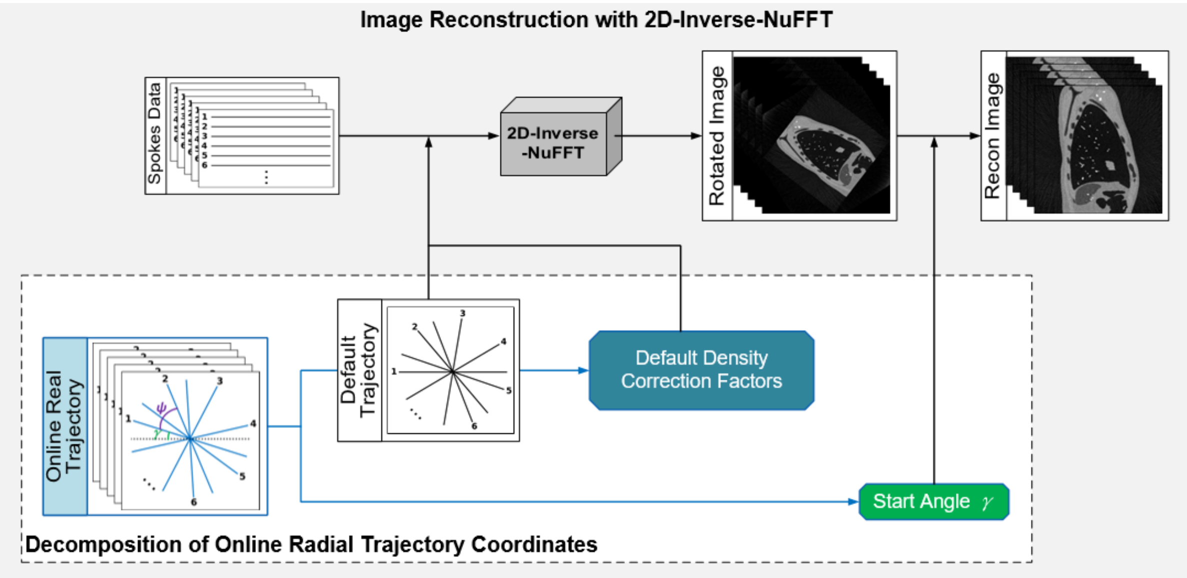

Figure 5.6: Decomposition of online radial trajectory coordinates and image reconstruction with 2D inverse NuFFT. This figure is adapted from material originally published in [166].

5.1.3 Radial dataset

All 25 simulated lung cancer patients were utilized to generate the radial dataset (see Section 3.4.1). For each patient, six original 2D+t radial cine-MR sequences were chosen from the 4D MRI phantom, covering sagittal, coronal and axial planes, with each plane containing two slices. The image matrices were produced with dimensions of 256×256-pixel arrays, featuring a spatial resolution of 1.5 mm $\times$ 1.5 mm. The SNR was configured to SNR $= 10$dB. All images were normalized by dividing them by their maximum magnitude values.

Following the methodology outlined in Chapter 3, 14 intra-frame motion patterns were employed to generate motion-corrupted frames from each original cine-MR sequence. For each pair of slices from the same plane, one slice was randomly assigned 7 out of the 14 patterns, while the other slice received the remaining 7. A total of 176 spokes were acquired for each frame, corresponding to $ns = 176$ shots, with $np = 340$ readout points sampled per spoke.

For each frame, motion-corrupted radial k-space spokes with a random starting angle γ for the first shot were acquired as input, and the corresponding final-position image served as the ground truth. The radial k-space spokes associated with the ground-truth image were also saved to facilitate the calculation of the sinogram loss $\mathscr{L}_{\text{sin}}$.

With ψ set to the golden angle $\psi_{\text{gold}} \approx 111.24°$, and two tiny golden angles, $\psi_5 \approx 111.24°$ and $\psi_{10} \approx 16.95°$, three radial datasets were generated for network training. Each dataset comprised 16800 (20 patients $\times$ 6 slices $\times$ 20 frames $\times$ 14 patterns $\times$ 1/2) input-output sample pairs for training (18 patients) and validation(2 patients), with the remaining 5 patients reserved for testing purposes.

5.1.4 Comparative Architectures and Implementation Details

To enable a comparative analysis against the TransSin-UNet, networks were also trained with architectures relying solely on a SinTE of $N = 12$ identical blocks and a 5-level UNet, using $\mathscr{L}_{\text{sin}}$ and $\mathscr{L}_{img}$, respectively. The first 4 levels of UNet matched those in the TransSin-UNet, except for the input channel number, as only the motion-corrupted image was fed into it.

It is worth noting that the configurations of these comparative architectures, such as the number of layers, were determined following the typical designs of Transformer and UNet models commonly used in the field [159, 167]. Given the critical importance of inference time in real-time motion management applications, a balance between performance and computational efficiency was pursued. Consequently, the TransSin-UNet in this study was designed with a more compact yet sufficiently deep architecture, incorporating an 8-layer Transformer Encoder and a 4-level UNet. Preliminary experiments were conducted and indicated that varying the number of layers had only a marginal impact on their performance. Fig. 5.7 shows a comparison of the training and validation loss curves across these different experimental conditions.

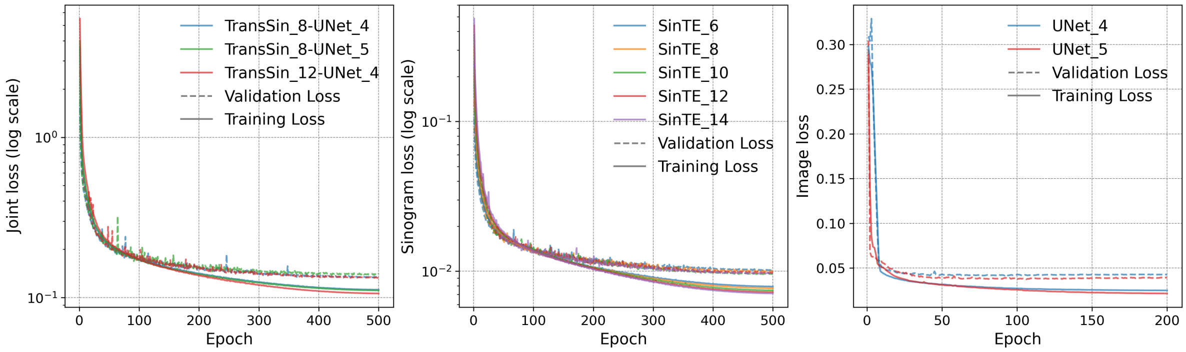

Figure 5.7: Training and validation loss curves for comparative architectures with varying numbers of layers. The notations "TransSin_n-UNet_m", "SinTE_n" and "UNet_m" refer to TransSin-UNet, SinTE and UNet models, respectively, where n indicates the number of SinTE blocks and m denotes the number of UNet layers. The dashed lines represent the validation loss, while solid lines correspond to the training loss. This figure was originally published as appendix material in [166].

The effectiveness of the models in image quality enhancement was quantitatively evaluated via a range of metrics, including structural similarity (SSIM), MAE, MSE or the normalized root mean squared error (NRMSE), which is normalized by the average Euclidean norm of the ground-truth image. The SSIM between two images x and y is formulated as [170]:

$$\mathrm{SSIM}(x, y) = \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right) \cdot \left(\frac{2\sigma_{x,y} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right) \tag{5.7}$$

where μ_x and μ_y are the mean values of x and y; σ_x^2 and σ_y^2 represent the variance; $\sigma_{x,y}$ is the covariance between x and y. $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$ are constants used to stabilize the division when the denominator is weak, with L being the dynamic range of the pixel-values. Typically $K_1 = 0.01$, $K_2 = 0.03$, and $L = 2^{\#\text{bits per pixel}} - 1$.

To quantitatively assess the target positioning accuracy of the models, GTV contours were generated for all sagittal frames containing tumors in the testing datasets, following the process detailed in Section 4.1.3. The DSC and average Hausdorff distance (HD_{avg}) were calculated to compare the GTV contour with the ground truth. The HD_{avg} is defined as:

$$\mathrm{HD}_{avg}(A, B) = \frac{1}{2}\left(\frac{1}{|A|}\sum_{a\in A}\min_{b\in B}\|a-b\| + \frac{1}{|B|}\sum_{b\in B}\min_{a\in A}\|b-a\|\right) \tag{5.8}$$

where $\|a - b\|$ is the Euclidean distance between points a and b. Consequently, a lower HD_{avg} signifies that the GTV contour is closer to the ground truth.

Inaccuracies or complete failures in GTV contouring with the optical-flow algorithm were observed in the radial testing datasets, as illustrated representatively in Fig. B.1 of the appendix. To alleviate this issue, a criterion was established for each comparison group:

$$\mathcal{A}\left(\mathbf{GTV}_{ref}\right) - \mathcal{A}\left(\mathbf{GTV}_{true}\right) \leq \quad 10\% \times \mathcal{A}\left(\mathbf{GTV}_{ref}\right) \tag{5.9}$$

where $\mathcal{A}$ represents the area calculator; $\mathbf{GTV}_{true}$ denotes the GTV contour obtained from the ground-truth final-position image via the optical-flow algorithm; and $\mathbf{GTV}_{ref}$ denotes the GTV contoured on the reference frame. This criterion is grounded on the assumption that tumor motion primarily follows a rigid pattern along the superior-inferior and anterior-posterior directions within the sagittal plane. Comparison groups with ground-truth GTVs that failed to fulfill this criterion were filtered out and excluded from the quantitative GTV positioning accuracy evaluation process.

To analyze whether there were statistically significant differences among (i) the performance of the three models, or (ii) the three datasets, Kruskal-Wallis tests were conducted with a significance level set at 0.01. If the Kruskal-Wallis test revealed a significant difference, a post-hoc Dunn test was performed to enable pairwise comparisons and further examine the specific variations.

NuFFT in TransSin-UNet was implemented using the torchkbnufft toolkit [171]. All the networks were developed with the PyTorch library [164], and were trained and tested on an NVIDIA A40 GPU with 48GB of memory, with a batch size of 16. The AdamW optimizer [172] was employed throughout the training process. For TransSin-UNet and SinTE, the learning rate was adjusted with a cosine annealing schedule [173], ranging from 10^{-4} to 5×10^{-6}; the training and validation loss curve approached stability after 500 epochs. For UNet, the learning rate started at 10^{-4} and was reduced by a factor of 0.8 if no improvement was observed in 12 epochs; the best validation loss was obtained at epoch 151. The weight parameter α in the joint loss function (Equation 5.2) was set to 10 to balance the two components.

5.2 Results

5.2.1 Inference time

The time consumed by the intra-frame motion compensation models is of vital importance in the application scenario of this study. Table 5.1 summarizes the average computational time of the models on the testing dataset. Notably, in the TransSin-UNet, the motion-corrupted image and the SinTE-corrected image are reconstructed in parallel before being fed into the UNet. Therefore, all compensation models involve one round of image reconstruction time, which remains constant compared to the conventional approach involving direct image reconstruction with motion-corrupted spokes in the reconstruction window. To ensure a fair comparison, data transfer and image reconstruction time costs were excluded from the analysis. Instead, the focus was placed solely on assessing the additional time required to compensate for the intra-frame motion.

It is demonstrated that all models can process one frame within a few milliseconds (ms), significantly shorter than the time span within the reconstruction window. Among them, the TransSin-UNet took an average of 4.87 ms. This outcome highlights the efficient performance of the TransSin-UNet, making it well-suited for real-time motion compensation applications.

Table 5.1: Inference time (mean $\pm$ std.) of intra-frame motion compensation models across the testing dataset, excluding data transfer and image reconstruction time. This table was originally published in [166].

	TransSin-UNet			UNet	SinTE
	Subnetwork: SinTE(N=8)	Subnetwork: UNet (4-level)	**Total**	(5-level)	(N=12)
Time per frame (ms)	3.14 ± 0.47	1.73 ± 0.28	$\mathbf{4.87 \pm 0.76}$	1.97 ± 0.40	4.28 ± 0.61

5.2.2 Performance Evaluation

Table 5.2 lists the quantitative testing results before and after intra-frame motion compensation with different models across the three datasets. Compared with other models, TransSin-UNet achieved the most substantial reduction in NRMSE and improvement in SSIM: on average, NRMSE was halved, and SSIM increased by 10.1%. The UNet was quantitatively less effective than TransSin-UNet, yet it was found to perform slightly better than SinTE across all investigated test subjects. The resulting p-values of Kruskal-Wallis tests and post-hoc Dunn tests within each dataset were consistently below 0.01, indicating statistically significant differences among the performances of the three models.

Table 5.2 also demonstrates minimal variations in results across different test sets. Further analysis to explore statistically significant differences among the three datasets was conducted using Kruskal-Wallis tests on metrics of MSE, MAE and SSIM, both before and after intra-frame motion compensation with each model. As shown in Table B.1 in the Appendix, no significant difference among the three datasets was observed, except for SinTE, which yielded significantly different results among the datasets in terms of SSIM. Post-hoc Dunn tests for SSIM values obtained with SinTE indicate significant differences between ψ_{10} and the other datasets. The specific p-values are provided in Table B.2 in the Appendix.

Table 5.2: Quantitative comparison of testing frames pre- (Motion-Corrupted) and post-intra-frame motion compensation. NRMSE and SSIM (Median [IQR]) are reported for each dataset; Mean values across all datasets are provided. The best results are highlighted in bold. This table was originally published in [166].

	NRMSE				SSIM (%)			
	ψ_{gold}	ψ_5	ψ_{10}	Mean	ψ_{gold}	ψ_5	ψ_{10}	Mean
Motion-Corrupted	0.147 [0.160]	0.146 [0.160]	0.146 [0.158]	0.188	77.6 [32.6]	77.3 [32.3]	77.5 [32.7]	72.9
TransSin-UNet	**0.094 [0.059]**	**0.092 [0.054]**	**0.091 [0.056]**	**0.092**	**82.2 [16.4]**	**82.1 [15.7]**	**82.2 [16.2]**	**83.0**
UNet	0.130 [0.119]	0.133 [0.122]	0.134 [0.124]	0.144	79.8 [20.3]	79.4 [19.8]	79.4 [20.5]	79.4
SinTE	0.150 [0.051]	0.150 [0.050]	0.148 [0.050]	0.159	70.8 [8.7]	70.7 [8.9]	71.6 [8.9]	69.6

Similar to the Cartesian experiments (in Section 4.2), intra-frame motion in the radial testing set was also categorized into three scenarios: *Static*, *Normal*, and *Extreme*. Fig. 5.8 compares the MSE of the testing frames across these motion scenarios before and after intra-frame motion compensation, while Fig. 5.9 depicts the GTV positioning accuracy of all sagittal frames containing tumors.

It is apparent from the results in Fig. 5.8 that, in the *Static* scenario, where the target remains stationary throughout the reconstruction window, the image before compensation is identical to the ground truth. TransSin-UNet and UNet exhibited comparable performance in this scenario, with the original data showing marginal changes after being processed by either of these two models. Notably, TransSin-UNet displayed slightly better stability, as indicated by its smaller interquartile range (IQR) in terms of image MSE. In contrast, SinTE appeared relatively weaker in maintaining image fidelity: the MSE exhibited noticeable increases, and numerous outliers were found in DSC, suggesting that the degradation of image quality could potentially impact the effectiveness of the applied optical-flow registration approach.

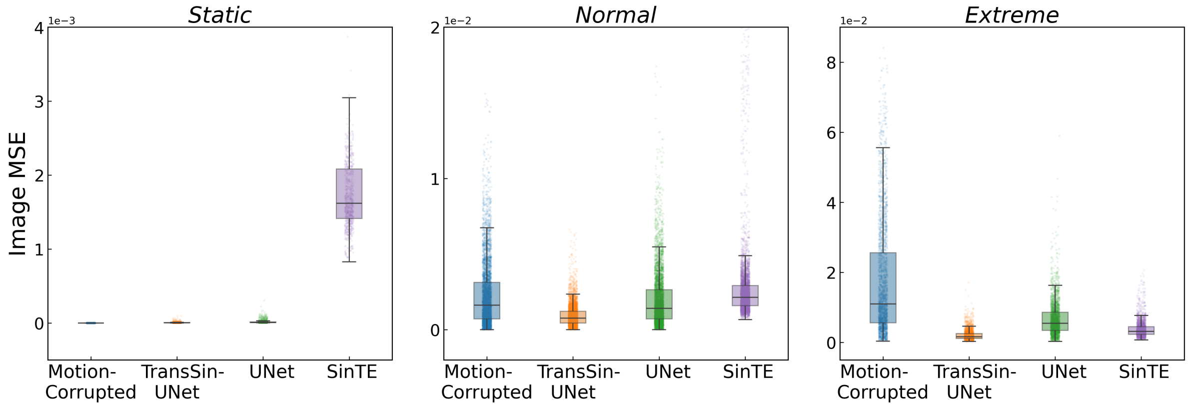

Figure 5.8: Box plot comparing the MSE of the testing frames pre- (Motion-Corrupted) and post- intra-frame motion compensation across three motion scenarios: *Static*, *Normal* and *Extreme*.The whiskers boundaries are based on 1.5 IQR. Note: in the *Static* scenario, the image before compensation is motion-corruption-free but is still labeled as "Motion-Corrupted" for convenience. Figure adapted from [166].

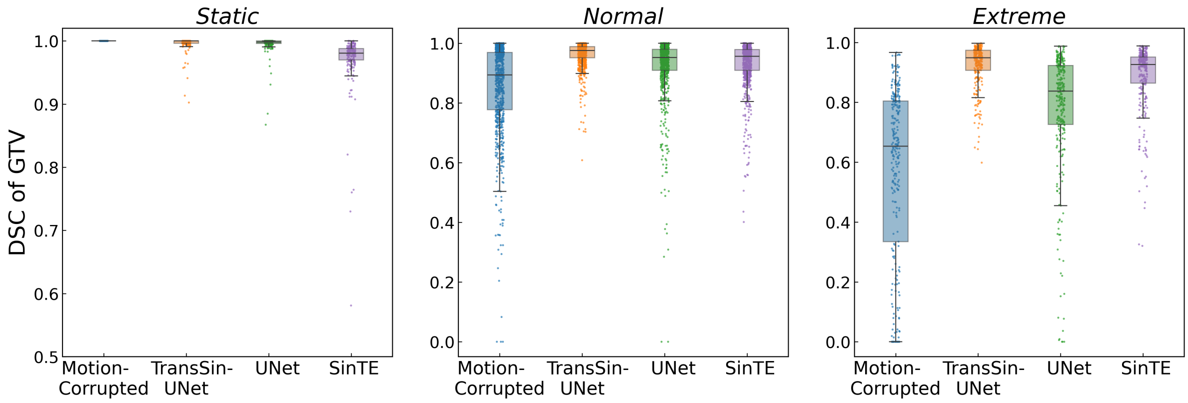

Figure 5.9: Box plot comparing target positioning errors pre- (Motion-Corrupted) and post- intra-frame motion compensation across three motion scenarios: *Static*, *Normal* and *Extreme*. DSC of GTV in all the sagittal testing frames containing tumors. The whiskers boundaries are based on 1.5 IQR. Note: in the *Static* scenario, the image before compensation is motion-corruption-free but is still labeled as "Motion-Corrupted" for convenience. Figure adapted from [166].

TransSin-UNet outperformed the other two models in both *Normal* and *Extreme* scenarios, effectively compensating for all kinds of intra-frame motion. This was evidenced by a remarkable reduction in image MSE and improvement in the median DSC of GTV: from 89.4% and 65.4% to 97.6% and 94.9% in *Normal* and *Extreme* scenarios, respectively. The UNet model surpassed SinTE in terms of image MSE among the *Normal* cases, while both achieved similar accuracy regarding GTV contour positioning.

Nevertheless, SinTE exhibited greater potential in eliminating *Extreme* intra-frame motion deterioration effects compared to UNet.

Since a tail of very low DSC values was observed in Fig. 5.9, a separate analysis was conducted on the outliers identified in the boxplot. Through individual inspections, it was determined that, apart from two cases listed in Table 5.3 attributed to sudden rapid motion of the target during the final stages of signal acquisition within the reconstruction window, all other DSC outliers of TransSin-UNet were a result of inaccuracies or even complete failures in the optical-flow algorithm. Despite filtering out comparison groups with incorrectly contoured ground-truth GTVs following Eq. 5.9, there were cases, such as "Wrong Case 1-3" and "Wrong Case 3" in Fig. B.1 of the Appendix, where the area of the ground-truth GTV met the criterion, but the optical-flow failed to contour the GTV on TransSin-UNet images or misplaced it entirely. Notably, these cases were deemed completely acceptable upon visual inspection, particularly in the *Static* scenario where the images and GTV contours of TransSin-UNet and UNet exhibited no perceptible differences from the ground truth.

Table 5.3: Outliers in TransSin-UNet performance attributed to sudden rapid motion of the target during the final stages of frame acquisition. This table was originally published in [166].

Patient ID	Scenario	Dataset	Initial/Final GTV centroid position	Extremum GTV centroid position	Time step of extremum	DSC (%)			
						Motion-Corrupted	TransSin-UNet	UNet	SinTE
2	*Normal*	ψ_{10}	178.2 mm	181.7 mm	136	77.8	83.1	69.4	78.6
2	*Normal*	ψ_5	178.2 mm	181.7 mm	136	77.5	85.6	62.3	78.2

The two cases in Table 5.3 originate from the same frame, experiencing identical intra-frame motion trajectories but differing in azimuthal radial profile increments. It can be observed that during the acquisition of the first 136 spokes, the GTV COM moved by 3.5 mm from its initial position and rapidly retracted back over the course of the remaining 40 spokes' acquisition (around 22.7% of the overall acquisition time). This indicates that the instantaneous velocity of the target significantly exceeded 72.6 mm/s, which should be regarded as an *Extreme* scenario. Nevertheless, TransSin-UNet obtained the highest DSC value among the models in these cases.

Fig. 5.10 presents representative sagittal and coronal frames from the testing patients breathing in *Normal* conditions, showcasing their imaging errors. These instances were chosen without bias towards selecting models exhibiting either favorable or unfavorable performance. The patient was inhaling and lung tumor was generally moving downwards during the acquisition of the sagittal frame. Conversely, the coronal frame was acquired during exhalation, with the tumor moving upwards. However, the position derived from the motion-corrupted image lagged behind the ground-truth final position of the tumor corresponding to the last shot within the reconstruction window. Substantial errors were appreciable around the edges of the moving structures in the motion-corrupted image, indicating evident imaging latency.

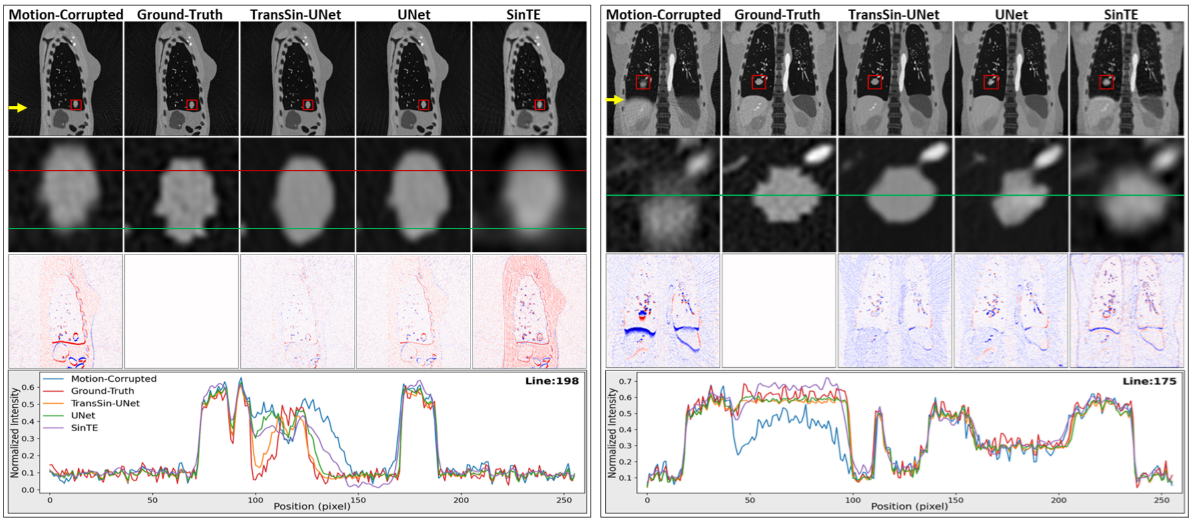

Figure 5.10: Image comparison pre- and post- intra-frame motion compensation. Exemplary sagittal (left) and coronal (right) frames are selected from the *Normal* motion cases of the testing patients. From top to bottom: Original image; Zoom-in image of the tumor taken from identical coordinate positions highlighted by the red box in the original image, with horizontal reference lines added to aid in perceiving the position; Error map, depicting the image difference with respect to the reference (calculated by subtracting the ground-truth); Intensity profile of the line along the yellow arrow in the original image. Figure reprinted from [166].

Among the models, TransSin-UNet demonstrated the closest tumor position and shape to the ground-truth, while UNet and SinTE tended to locate the tumor between the motion-corrupted and the ground-truth, offering only a partial compensation. By employing TransSin-UNet, the error map, as well as the intensity profile along an exemplary line in the motion-corrupted image, was effectively corrected, achieving the highest agreement with the ground-truth. While SinTE displayed slightly better performance over the UNet model in preserving the tumor shape, particularly in the coronal example, it exhibited a tendency to produce more blur in the output images.

A few failure cases of the UNet were observed. Fig. 5.11 shows one example where the errors in the UNet-generated image exceed those in the motion-corrupted input. This indicates that the primary focus of the UNet lies in enhancing the clarity of the blurred structural edges by refining intensity, but it may have misinterpreted the anatomical structures corresponding to these areas. In contrast, SinTE and TransSin-UNet prioritize target positioning, and TransSin-UNet also demonstrates a high capacity for deblurring.

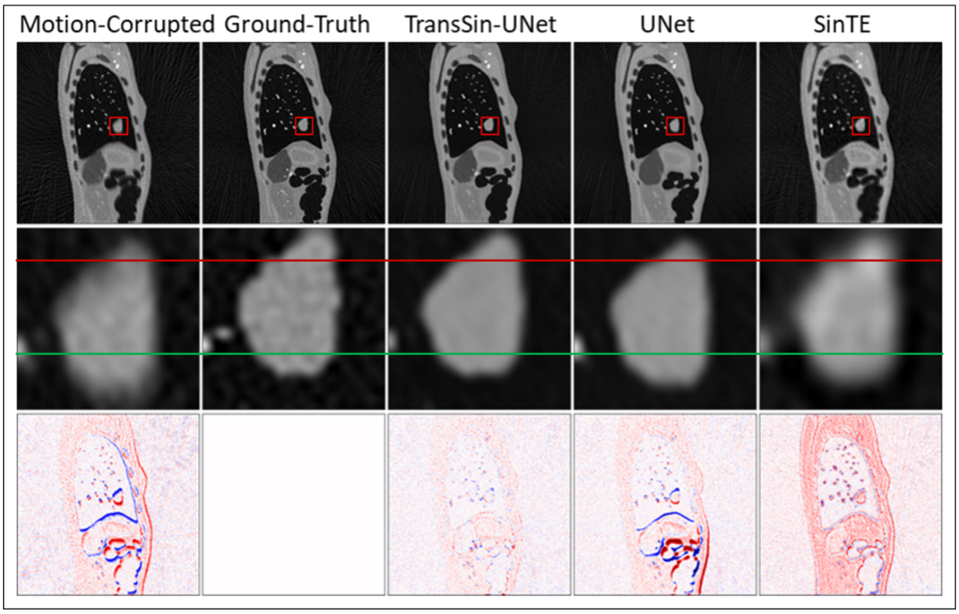

Figure 5.11: Representative example where UNet failed to provide effective compensation. From top to bottom: Original image; Zoomed-in image of the tumor taken from the same coordinate positions highlighted by the red box in the original image, with horizontal reference lines added to aid in perceiving the position; Error map, depicting the image difference relative to the reference (calculated by subtracting the ground-truth).

An additional target positioning accuracy evaluation was conducted, focusing on sagittal test subjects experiencing *Normal* motion conditions from all the datasets. These subjects were categorized into three subgroups based on the GTV COM shift of the motion-corrupted image from the ground-truth: *Small* (COM shift $\leq$ 1.5 mm), *Medium* (1.5 mm $<$ COM shift $\leq$ 4.5 mm) and *Large* (COM shift $>$ 4.5 mm). In Table 5.4, the GTV structures before and after intra-frame motion compensation were compared to the ground-truth via the DSC and HD_{avg}. The results from the motion-corrupted images underscored the importance of implementing compensation in radial cine-MR, given that the intra-frame motion in *Medium* and *Large* groups yielded median DSC values as low as 78.3% and 62.2%, respectively.

Table 5.4: GTV positioning accuracy of test subjects moving in a *Normal* scenario, pre- (Motion-Corrupted) and post- intra-frame motion compensation. Median [IQR] of DSC and HD_{avg} are reported for the *Small*, *Medium* and *Large* subgroups, respectively. Mean values across all subgroups are provided. The best results are highlighted in bold. This table was originally published in [166].

	DSC (%)				HD_{avg} (mm)			
	Small	*Medium*	*Large*	Mean	*Small*	*Medium*	*Large*	Mean
Motion-Corrupted	96.1 [6.8]	78.3 [9.4]	62.2 [14.0]	85.1	0.039 [0.069]	0.341 [0.266]	1.111 [0.650]	0.389
TransSin-UNet	**98.4 [2.1]**	**95.8 [5.0]**	**94.1 [5.0]**	**96.2**	**0.016 [0.021]**	**0.042 [0.051]**	**0.062 [0.049]**	**0.047**
UNet	97.1 [4.4]	92.4 [7.3]	84.6 [26.0]	92.0	0.028 [0.045]	0.080 [0.092]	0.323 [0.761]	0.192
SinTE	97.2 [2.8]	90.9 [6.1]	86.7 [10.5]	92.9	0.028 [0.028]	0.094 [0.074]	0.163 [0.213]	0.125

Within each group, TransSin-UNet emerged as the most powerful model in reducing the target positioning errors. It improved the median DSC by 17.5% in the *Medium* group and by 31.9% in the *Large* group, while reducing the median HD_{avg} by approximately 59%, 88%, and 94% from the initial values of 0.039, 0.341, and 1.111 mm in the *Small*, *Medium*, and *Large* groups, respectively. Moreover, the decrease in IQR demonstrated the stability of its performance.

Less effective than TransSin-UNet, UNet and SinTE delivered similar outcomes in *Small* and *Medium* groups: they drove the median DSC to over 97.1% in the former group and over 90.9% in the latter; additionally, they brought a decrease in the median HD_{avg} to 0.028 mm in the *Small* and below 0.094 mm in the *Medium* group. Nevertheless, SinTE demonstrated superior performance compared to UNet in the *Large* group, increasing DSC by 24.5% and reducing HD_{avg} by 0.948 mm (around 85% of the initial value).

Targets and OARs have varying extents of structural deformation during MRgRT. Besides assessing the models' accuracy on small mobile targets, their performance concerning large target deformation was further evaluated. This assessment can be conducted by examining the axial slices where specific inter-slice motion can be interpreted as deformation in 2D, as shown in Fig. 5.12 and Fig. 5.13.

Comparing the motion-corrupted image with the ground truth, it can be inferred that significant intra-frame deformation likely occurred in the lung, the liver, and around the heart within the time span of the reconstruction window.

In Fig. 5.12, the liver area derived from the motion-corrupted image appeared notably diminished and the tumor was nearly imperceptible, reflecting substantial geometric tracking errors of MR-guidance. All three compensation models were able to successfully detect the presence of the tumor in the slice at the end of the frame acquisition. Particularly noteworthy, TransSin-UNet closely mirrored the true shapes of all the anatomical structures. Although UNet tended to produce a sharper tumor image compared with SinTE, its performance in restoring the shapes of the liver and blood in the cardiac region was inferior; conversely, SinTE excelled in capturing large structural changes but was less effective in preventing image blurring.

In Fig. 5.13, the blood within the cardiac image underwent a notable shape transformation during the time span of the reconstruction window. The shape distortion in the motion-corrupted image was effectively rectified, and was aligned closely with the ground-truth using TransSin-UNet. Moreover, SinTE also performed well in this case, despite yielding blurred edges. This suggests that both models have gained the ability of extracting the latent structural information of the final-position image from the chronologically last few spokes in the window. However, the UNet model primarily focused on deblurring the motion-corrupted image and enhancing the details, but was incapable of delineating a correct structural shape.

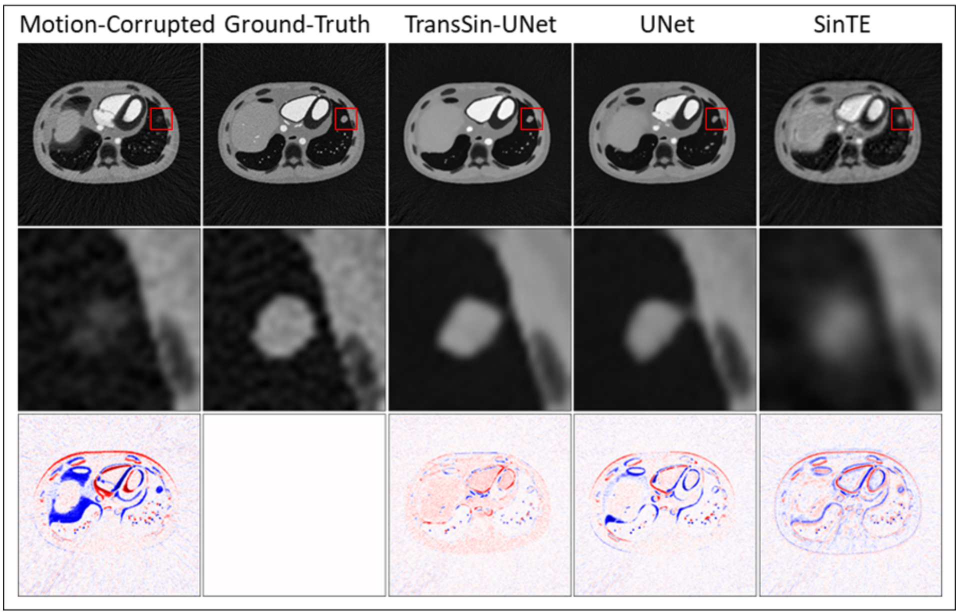

Figure 5.12: Deforming target evaluation: Representative axial frames under *Normal* motion conditions. From top to bottom: Original image; Zoomed-in view of the tumor, highlighted by the red box in the original image; Error map showing the difference with respect to the reference, calculated by subtracting the ground truth. This figure is adapted from material originally published in [166].

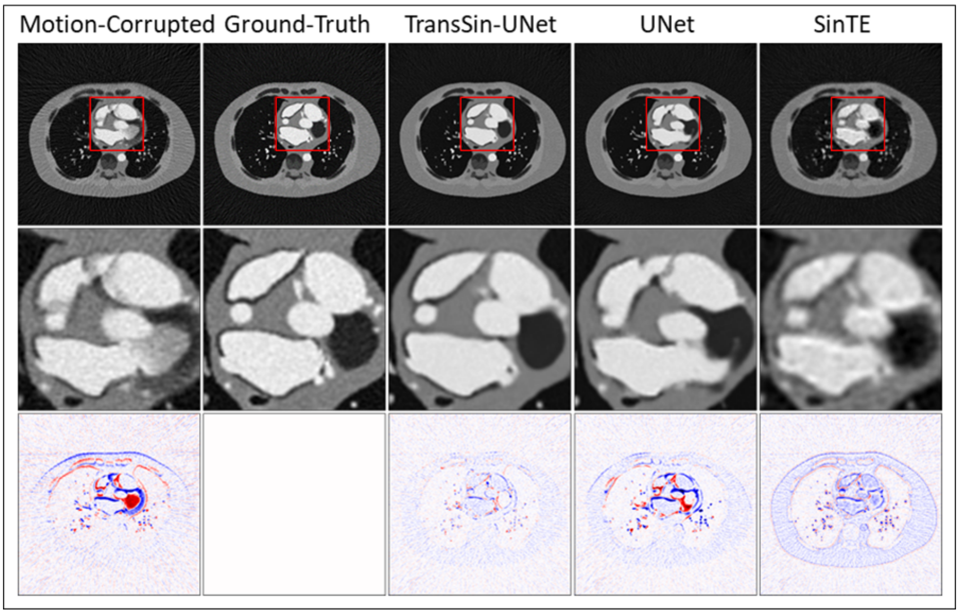

Figure 5.13: Deforming target evaluation: Representative axial frames under *Normal* motion conditions. From top to bottom: Original image; Zoomed-in view of the cardiac region, highlighted by the red box in the original image; Error map illustrating the difference relative to the reference, computed by subtracting the ground truth. This figure is adapted from material originally published in [166].

5.3 Discussion

This chapter investigates the feasibility of reducing imaging errors in radial cine-MRI by implementing deep learning-based intra-frame motion compensation techniques.

Instead of compromising the reconstruction window width, this study hypothesized that the radial spokes positioned earlier in the window could still provide effective information for final-position image reasoning. TransSin-UNet was designed to operate in both the projection and spatial domains. The SinTE subnetwork learns the long-distance spatial-temporal dependencies between the sinogram representations of the spokes, calibrating them to align with those of the ground-truth final-position image. Furthermore, the UNet subnetwork is responsible for fine-tuning the local details, enabling pixel-level enhancement in the spatial domain.

By mapping the spoke data from the frequency domain to the projection domain based on the Fourier projection-slice theorem, the spoke signal values were converted to a scale range comparable to the original image intensity values. This technique successfully addresses the issue of large magnitude disparities inherent in k-space data, thereby improving gradient behavior for the non-linear activation functions in the transformer encoder and contributing to a more stable training process.

The relative spatial and temporal positional encodings are unified in SinTE based on the chronological order of acquired spokes, rendering the model agnostic to a specific radial sampling trajectory. Consequently, for a given azimuthal radial profile increment ψ, the model can accommodate trajectories with any arbitrary initial spoke angle.

Inference time is a critical factor for the practical deployment of real-time motion management. Given that the output of the SinTE subnetwork is from the same domain as the input, unlike existing work on k-space transformer that retains both the encoder and decoder structures [108, 131], TransSin-UNet only incorporates the encoder component of the transformer. This design choice circumvents the potential time consumption associated with an auto-regressive decoder. Additionally, the decomposition of online radial trajectories eliminates the need for the time-consuming online DCF computation. To date, in clinical MR-Linac systems, the frame rate of radial cine-MRI is 8 Hz (i.e. 125 ms/frame) [32], which is related to the stride of the sliding window shown in Fig. 5.1. As shown in Table 5.1, under the GPU configuration utilized in this study, TransSin-UNet introduced only an additional 4.8 ms per frame compared to the conventional approach—an overhead that is negligible when considering the duration of the reconstruction window.

In Section 5.1.2, TransSin-UNet and architectures relying solely on UNet or SinTE were compared through extensive quantitative and qualitative evaluations.

Compared to SinTE, UNet displayed notable advantages in edge sharpening and image deblurring (refer to Fig. 5.10 $\sim$ Fig. 5.13); quantitative results depicted in Fig. 5.8 and Fig. 5.9 demonstrated that the MSE approached zero in the *Static* scenario and was lower in the *Normal* scenario. Nevertheless, the UNet model showed limited effectiveness in compensating for larger anatomical changes. It failed in several cases, as representatively shown in Fig. 5.11, and performed the worst in the *Extreme* scenario (Fig. 5.8 and Fig. 5.9) among the three models. The median DSC in the *Large* group (Table 5.4) was 2.1% and 9.5% lower than those of SinTE and TransSin-UNet, respectively. The previous chapter highlighted the effectiveness of UNet in Cartesian experiments, emphasizing its particular adeptness at identifying information associated with specific frequency ranges and efficiently alleviating imaging errors induced by varying levels of intra-frame motion. However, in a radial MR trajectory, each spoke crosses the origin and uniformly spans both high and low frequencies of the k-space. As clearly shown in Fig. 5.12 and Fig. 5.13, larger intra-frame anatomical changes pose significant challenges for UNet in fully leveraging its advantages and generating precise final-position structural shapes.

SinTE, on the other hand, utilizes positional encoding to directly incorporate the relative spatial and temporal information of the spokes. The results from the *Extreme* scenario (Fig. 5.8 and Fig. 5.9), the *Large* group (Table 5.4), and the deforming target (Fig. 5.12 and Fig. 5.13) demonstrated SinTE's capability to capture relatively large anatomical changes. However, SinTE was found to be less sensitive to smaller changes, which can be attributed to its operation in the image's projection domain.

As expected, TransSin-UNet integrates the strengths of both subnetworks, showcasing significantly superior image quality and enhanced accuracy in target positioning. Across all comparisons, TransSin-UNet consistently outperformed UNet and SinTE, irrespective of the motion trajectories or amplitudes, with metrics assessing image disparities achieving optimal values. In Table 5.4, the mean DSC of GTV in the investigated testing cases significantly improved, rising from 85.1% to 96.2%, while the median HD_{avg} in the *Large* group was notably reduced by 94%. Additionally, the decrease in IQR further emphasizes the stability of the model. The final-position images of subjects experiencing considerable intra-frame deformations were precisely derived from the motion-corrupted radial spokes. These findings underscore the efficacy of TransSin-UNet in mitigating radial cine-MR imaging errors by effectively accounting for the target motion within the reconstruction window.

Table 5.2 compares the quantitative outcomes across three distinct datasets characterized by varying angular increments: ψ_{gold}, ψ_5 and ψ_{10}, revealing minimal variation among them. Specifically, motion-corrupted images acquired with different profile ordering schemes demonstrated similar sensitivity to the intra-frame motion, as evidenced by the non significant p-value in the Kruskal-Wallis tests (Table B.1). This behavior is explicable considering the incoherence properties of these trajectories: unlike the $linear$ radial trajectory where temporally close spokes are also spatially close, the *(tiny) golden angle* acquisition may interleave the newly acquired spokes with the previously acquired ones. As a result, in a specific time step interval, motion-related data variations do not concentrate in a specific high- or low- frequency region but rather uniformly disperse throughout the entire k-space. Although the input tokens changed with respect to ψ, SinTE was able to yield comparable regression results in terms of pixel-wise metrics by establishing their spatial-temporal interactions. Nonetheless, structure discrepancies introduced by it can be statistically significant (see Table B.2). Moreover, when taking similar motion-corrupted images as input, UNet operates in the spatial domain, and no significant differences were observed in the output. As evidenced by p-values larger than the significance level in all evaluating metrics, TransSin-UNet demonstrated strong robustness to the azimuthal profile increment of the radial trajectories.

5.4 Conclusions

This chapter proposes reducing errors in radial cine-MR imaging by implementing intra-frame motion compensation techniques. A novel network (TransSin-UNet) was designed and successfully trained with datasets characterized by varying azimuthal k-space radial profile increments in lung cancer cases. The model effectively derived the final-position image of the subject corresponding to the end time of the reconstruction window. Results showed that TransSin-UNet outperformed architectures relying solely on UNet or SinTE across all the investigated comparative experiments, leading to significant improvements in image quality and target positioning accuracy. In conclusion, TransSin-UNet demonstrated great potential in continuously compensating for target motion within the sliding window of radial cine-MR acquisition, thereby enhancing real-time imaging accuracy for MRgRT.

Chapter 6

SUMMARY AND OUTLOOK

6.1 Summary

Motion-related imaging errors have been recognized as the primary contributor to overall loop latency of MRgRT, leading to residual geometric tracking errors and, consequently, affecting the effectiveness of active intra-fractional motion management. To the best of the author's knowledge, this study represents the first attempt to investigate the feasibility of mitigating motion-related errors in real-time MR imaging by implementing deep learning-based intra-frame motion compensation techniques.

Since MRI raw signal data are acquired in the frequency domain, a dedicated procedure was developed in this thesis to investigate the dynamic MR imaging behavior. The motion-dependent k-space sampling simulation revealed that, as the acquisition of a single cine-MR frame occurs on the same time scale as physiological motion, the resulting k-space incorporates signals from the target at varying positions, leading to effective motion-induced errors in the spatial domain. For both linearly and fully acquired Cartesian readout, as well as radial readout trajectories, intra-frame motion resulted in an imaging latency of approximately 50% of the time span of the sampled k-space data used for image reconstruction. This underscored the practical value of implementing intra-frame motion compensation, particularly in cases of rapid breathing or for anatomical structures influenced by the cardiac motion. An ill-posed inverse problem was then formulated to recover the implicit real-time final-position image, corresponding to the end of the frame acquisition, from the motion-corrupted image or k-space.

To address this issue, data-driven deep learning-based approaches have emerged as the most prominent solution, leading to the development of a methodology for intra-frame motion dataset creation and augmentation, with the simulation code serving as the data generator and focusing on rapid anatomical changes. Based on coarse-to-fine grid-scale representation of patient-specific motion data, 25 4D MRI digital anthropomorphic phantoms were generated to model lung cancer patients, and a dedicated intra-frame motion model was constructed with a piecewise linear approximation between consecutive control points. Additionally, a motion pattern perturbation scheme was introduced to comprehensively explore the potential anatomical structure positions and enhance the diversity of intra-frame motion trajectories. This framework establishes a foundation for generating and augmenting intra-frame motion datasets from physical experimental data, supporting future deep learning applications in clinical practice.

The *in silico* proof-of-concept study for Cartesian cine-MRI was presented in Chapter 4. The UNet models were successfully trained to estimate the final-position images at the end of acquisition from the motion-corrupted input, demonstrating high effectiveness in intra-frame motion compensation. Quantitatively, for the testing dataset analyzed in GTV contouring, the median DSC increased from 89% to 97%, while the HD_{95} decreased from 4.1 mm to 1.4 mm. Additionally, geometric errors caused by intra-frame anatomical deformations in certain regions were successfully corrected, in terms of both target shape and position.

The network's versatility was demonstrated in the undersampled Cartesian MRI experiment, where it simultaneously performed undersampling-based acceleration and intra-frame motion compensation, effectively mitigating both aliasing artifacts and residual geometric tracking errors. Furthermore, saliency maps of the motion-corrupted input highlighted the major contribution of later-acquired k-space data to model inference and, correspondingly in the spatial domain, the edges of the moving anatomical structures at their final positions. These behaviors are particularly relevant in addressing concerns regarding the feasibility and reliability of deep learning approaches for clinical implementation.

The *in silico* proof-of-concept study for radial cine-MRI was presented in Chapter 5. It was noticed that while radial sampling allows for nearly arbitrary frame rates with sliding window reconstruction, imaging latency is independent of the frame rate and instead depends on the temporal coverage of spokes within the reconstruction window. Compared to Cartesian sampling trajectories, radial sampling exhibits distinct characteristics. Firstly, each spoke passes through the origin of k-space and spans both high and low frequencies. Secondly, with the sliding window method, the first spoke within a specific window can be positioned at an arbitrary angle, resulting in unique trajectory coordinates for each frame. Moreover, *(tiny) golden angle* acquisitions may interleave newly acquired spokes with previously acquired ones, leading to cases where temporally close spokes are not necessarily spatially close. Consequently, the interactions among the spokes were expected to be modeled with consideration given to both spatial and temporal adjacency.

Instead of compromising the window width, a novel network, TransSin-UNet, was proposed to accommodate the nature of radial k-space readout trajectories. The model operates in both the projection and spatial domains, with a joint loss function defined as a weighted linear combination of respective losses from each domain. Specifically, a transformer encoder with its attention mechanism was employed to model the long-range dependencies between the spokes, aligning them with the ground truth, followed by pixel-level fine-tuning in the spatial domain with a UNet.

The reason for operating in the projection domain rather than the frequency domain is related to another important consideration: the power spectrum characteristics of medical images, where the center of k-space exhibits significantly higher energy than the peripheral regions, and the values along each spoke span a wide range of magnitudes. Directly using these values as input may lead to poorly suited gradients for the non-linear activation functions in the transformer encoder, potentially impeding convergence. To address this, the model converts each spoke into its sinogram representation, which corresponds to a projection of the MR image along that spoke, as described by the Fourier projection-slice theorem. This transformation reduces the dominance of central k-space values, ensuring a more balanced magnitude distribution across all token dimensions and thereby facilitating stable processing in the transformer encoder.

TransSin-UNet combined the advantages of transformer encoders in capturing relatively large intra-frame anatomical changes and UNet in edge sharpening. It consistently outperformed architectures relying solely on transformer encoders or UNets across all comparative evaluations, leading to a noticeable enhancement in image quality and target positioning accuracy. The NRMSE decreased by 50% from an initial average of 0.188, while the mean DSC of GTV increased from 85.1% to 96.2% in the investigated testing cases. Final-position images of anatomical structures undergoing substantial intra-frame deformations were accurately derived from the motion-corrupted input. Moreover, TransSin-UNet maintained robust performance across datasets with varying azimuthal radial profile increments.

The inference time is critical to the research problem addressed in this thesis, which directly determines the feasibility of the techniques in practical applications. This aspect was a key focus of this thesis. TransSin-UNet was designed to incorporate only the encoder component of the transformer to circumvent the potential extensive computation time associated with an auto-regressive decoder. Additionally, the online trajectory coordinates of each frame were decomposed into the unified default trajectory with the frame-specific starting angle, which conserved storage space and eliminated the need for time-consuming online calculation of DCFs for reconstruction with NuFFT. Compared to the conventional approach involving direct image reconstruction with motion-corrupted k-space data, the models required only a few additional milliseconds to complete the motion compensation. This inference time is negligible when compared to the frame acquisition time or the reconstruction window duration.

In conclusion, this thesis introduced a novel concept of motion-related imaging errors in MRgRT and proposed their reduction through deep learning-based intra-frame motion compensation techniques. A motion-dependent k-space acquisition simulation procedure was developed, and a methodology for intra-frame motion dataset creation and augmentation was introduced, with a primary focus on rapid anatomical variations. Proof-of-concept studies were conducted on both Cartesian and radial cine-MRI acquisitions, respectively. A novel model, TransSin-UNet, was proposed, specifically tailored for radial sampling. An extensive *in silico* feasibility analysis was performed, encompassing evaluations of image quality and target positioning accuracy, model comparison, and studies on versatility, robustness, and interpretability to assess the proposed approaches. The results highlight the significant potential of these methods for continuous intra-frame motion compensation in clinical settings, improving the accuracy of real-time MRI motion monitoring and further advancing intra-fractional motion management in MRgRT.

6.2 Outlook

Experimental validation of this study in clinical settings constitutes a crucial next step and is currently in progress. Nonetheless, obtaining paired motion-corrupted and ground-truth final-position images from the MR-Linac remains challenging, as detailed in Section 3.3. To address this limitation, this study proposes an approach termed *frame-merging* for validation with real clinical data, building upon the presented intra-frame motion dataset creation method.

Based on the findings of this work, patients undergoing breath-hold or very shallow breathing exhibit negligible intra-frame motion, allowing the acquired MR frames in these cases to be considered free of motion corruption. Images extracted from these stages, corresponding to different inhale/exhale amplitudes, are referred to as stopping-point frames. Patients or volunteers can be instructed to hold their breath at the middle or end of inhale or exhale phase to acquire these frames.

In the *frame-merging* method, the MR-Linac frame acquisition time is assumed to be increased by a factor of N, achieved by merging $N-1$ consecutive frames preceding the stopping-point frame into a single frame. Following the motion-dependent sampling process, these N frames serve as the temporal images, with intra-frame motion modeled as a function of their relative positions. Interpolation between these frames can be applied to refine the motion trajectory. The merging process is performed in k-space by extracting the corresponding components from the temporal frames and incorporating them into the motion-corrupted k-space array, with the stopping-point frame representing the last-shot position. The final merged image is treated as the motion-corrupted image, while the stopping-point frame is considered the ground truth. Furthermore, the presented motion pattern perturbation scheme enables dataset augmentation for training purposes.

Currently, the motion-dependent signal acquisition simulation procedure has been validated only through theoretical analysis and comparison with existing literature.

Further verification through imaging latency experiments on MR-Linacs can be conducted [32, 33, 88]. Other k-space trajectories, such as spiral, and MRI acceleration techniques, including partial Fourier and parallel imaging methods, warrant further investigation. The effects of these techniques have yet to be quantified, particularly for acceleration methods that involve sharing k-space data across different coil images, which may impact the extent of motion-related imaging errors.

One limitation of this study lies in the requirement to segment the GTV for quantitative validation. The extent to which the GTV positioning accuracy results were influenced by optical flow-based segmentation has yet to be precisely determined. Future investigations should include a supplementary uncertainty analysis specifically tailored to this or consider alternative and potentially more reliable contouring strategies, such as foundation models recently proposed for medical imaging tasks [174].

Given the exploratory nature of this proof-of-concept study, extensive hyperparameter tuning for the network was not conducted. Subsequent efforts necessitate a comprehensive hyper-parameter searching grounded in the real clinical data to identify the optimal network configurations. Exploring other multi-loss weighting approaches for TransSin-UNet holds particular promise [175, 176]. Additionally, while the spatial and temporal positional encoding in TransSin-UNet is unified and based on the chronological order of acquired spokes, other methods or architectural variants that factorize the spatial and temporal dimensions of the input tokens [177] could be explored.

The undersampled Cartesian MRI experiment demonstrated the versatility of the UNet model, as it simultaneously reduced image noise, aliasing artifacts and motion-related imaging errors. One of the next potential steps is to further explore the combination of intra-frame motion compensation and other tasks in MR imaging or MRgRT, such as geometric distortion correction [178,179], synthetic CT generation [180,181]. Given UNet's strong capabilities in image segmentation, both it and TransSin-UNet could be trained to directly generate segmented tumors or OARs as needed. Transfer learning, such as patient-specific adaptation [114], could be explored. Further investigation into the interpretability of the network and its generalization to out-of-distribution (OOD) data would also be valuable.

Further research is warranted to explore the potential benefits of integrating the proposed approach with other advanced techniques such as motion prediction [101]. Literature findings suggest that the efficacy of motion prediction algorithms improves as the forecasted time span decreases [182, 183]. As previously discussed, within the total loop latency of the MR-Linac, the contribution of MLC-related delays is assumed to be insignificant compared to the substantially greater latency introduced by MR imaging [32]. The approaches developed in this study effectively account for imaging latency, thereby significantly shortening the required prediction time span—an aspect that presents a promising opportunity for improving the accuracy of subsequent prediction algorithms. Moreover, recent studies indicate that predicting 2D tumor motion from cine-MRI frames is considerably more challenging than 1D centroid-based tumor motion prediction [104]. The proposed intra-frame motion compensation models operate in 2D, providing effective latency correction for anatomical variations in both position and shape, which potentially contributing to improved 2D motion prediction.

An extension of the proposed method is envisioned for closely related applications, such as real-time 4D-MRI and MR-integrated proton therapy (MRiPT) [184].

Signal acquisition in 3D generally takes longer than in 2D, and intra-frame motion compensation could bypass the consequent issue of inadequate temporal resolution in MR imaging. Therefore, generalizing this technique from 2D+t cine-MR to 3D+t represents a promising avenue for future research [185].

Proton therapy, and more broadly particle therapy [186], holds promise for achieving superior dose conformity compared to X-ray therapy due to the finite particle range and the presence of the Bragg peak [187]. However, particle therapy is more susceptible to uncertainties encountered in clinical workflows, with range uncertainty [188] being a major concern. Real-time MRI guidance in particle therapy represents a promising advancement for enhancing treatment delivery precision through improved motion monitoring and management [70]. In this context, motion-related imaging errors become increasingly critical. The intra-frame motion compensation strategy proposed in this work could therefore provide substantial benefits, making it a potential direction for future development.

Appendix A

Proof of the Translational and Rotational Properties of the Fourier Transform

A.1 Proof of the translational property of the Fourier transform

The translational property of the Fourier transform states that a translation in spatial domain results in a linear phase shift in the frequency domain. Given an image $f(x, y)$ and consider translating it by $(\Delta x, \Delta y)$ in the spatial domain. The translated image can be expressed as:

$$\hat{f}(x, y) = f(x - \Delta x, y - \Delta y) \tag{A.1}$$

The Fourier transform of this translated image, denoted by $\hat{F}(u, v)$, is given by:

$$\hat{F}(u, v) = \iint_{-\infty}^{\infty} f(x - \Delta x, y - \Delta y) \exp\left[-j2\pi(ux + vy)\right] \, dx \, dy \tag{A.2}$$

where, u and v are the frequency components in the Fourier domain, and i is the imaginary unit. By Substituting $x = \hat{x} + \Delta x$ and $y = \hat{y} + \Delta y$ into the equation, the Fourier transform of the translated image becomes:

$$\hat{F}(u, v) = \exp\left[-j2\pi(u\Delta x + v\Delta y)\right] \iint_{-\infty}^{\infty} f(\hat{x}, \hat{y}) \exp\left[-j2\pi(u\hat{x} + v\hat{y})\right] \, d\hat{x} \, d\hat{y} \tag{A.3}$$

The integral on the right is simply the Fourier Transform of the original image, denoted by $F(u, v)$. Thus, we have:

$$\hat{F}(u, v) = F(u, v) \exp\left[-j2\pi(u\Delta x + v\Delta y)\right] \tag{A.4}$$

This result demonstrates that shifting the image by Δx in the x-direction and by Δy in the y-direction introduces a phase shift of $2\pi(u\Delta x + v\Delta y)$ in the Fourier domain, while the magnitude of the Fourier Transform remains the same.

A.2 Proof of the rotational property of the Fourier transform

The rotational property of the Fourier transform states that a rotation in the spatial domain corresponds to a rotation by the same angle in the frequency domain. Specifically, consider rotating the image $f(x, y)$ by an angle θ around the origin. The rotated coordinates (x', y') can be expressed as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{A.5}$$

The Fourier transform $G(u', v')$ of the rotated image $g(x', y')$ is given by:

$$G(u', v') = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x', y') \exp\left[-j2\pi(u'x' + v'y')\right] \, dx' \, dy' \tag{A.6}$$

Substituting x' and y' with x and y, the Jacobian determinant of the transformation is 1, indicating that the transformation preserves area, and noting that $g(x', y') = f(x, y)$, the integral becomes:

$$G(u', v') = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) e^{-j2\pi[u'(x\cos\theta - y\sin\theta) + v'(x\sin\theta + y\cos\theta)]} \, dx \, dy \tag{A.7}$$

Simplifying the exponent:

$$G(u', v') = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) e^{-j2\pi[(u'\cos\theta + v'\sin\theta)x + (-u'\sin\theta + v'\cos\theta)y]} \, dx \, dy \tag{A.8}$$

This can be rewritten as:

$$G(u', v') = F(u, v) \tag{A.9}$$

where

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} u' \\ v' \end{bmatrix} \tag{A.10}$$

Eq. A.9 shows that the Fourier transform of the rotated image is the Fourier transform of the original image, but evaluated at the rotated coordinates (u', v').

Appendix B

Supporting Information

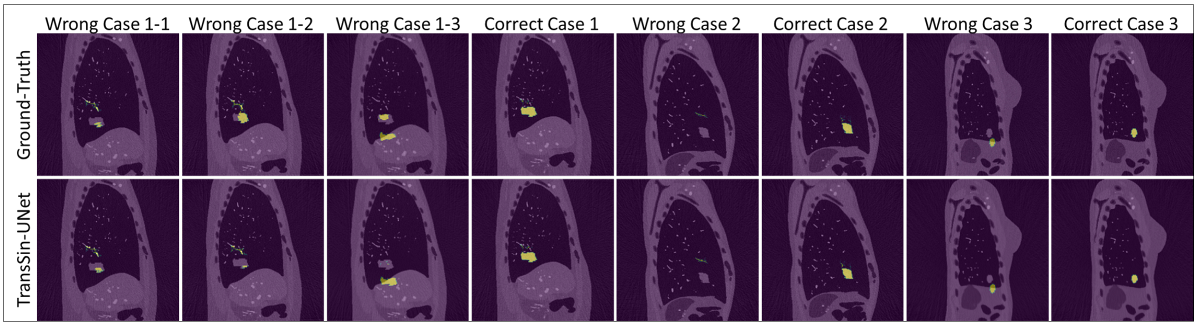

Figure B.1: Representative cases of inaccuracies or complete failures observed in the optical-flow algorithm for GTV contouring. The GTV segmentations generated by the optical-flow algorithm are highlighted in yellow on both the ground-truth and TransSin-UNet output images. This figure was originally published as appendix material in [166].

Table B.1: P-values obtained from the Kruskal-Wallis test for comparing differences among the three datasets (ψ_{gold}, ψ_5, and ψ_{10}). Results from metrics of MSE, MAE and SSIM pre- (Motion-Corrupted) and post- intra-frame motion compensation with different models are presented. Statistically significant values (p-value < 0.01) are indicated by an asterisk. This table was originally published as appendix material in [166].

	Motion-Corrupted	TransSin-UNet	UNet	SinTE
MSE	0.99	0.014	0.15	0.075
MAE	0.93	0.066	0.49	0.47
SSIM	0.81	0.41	0.58	8.5E-7*

Table B.2: P-values obtained from the post-hoc Dunn test for pairwise dataset comparisons using the SSIM metric after intra-frame motion compensation with SinTE. Statistically significant values (p-value < 0.01) are indicated by an asterisk. This table was originally published as appendix material in [166].

	ψ_{gold} and ψ_5	ψ_{gold} and ψ_{10}	ψ_5 and ψ_{10}
SSIM (SinTE)	0.48	5.2E-6*	6.3E-6*

Bibliography

[1] World Health Organization. Cancer overview, 2024. Accessed: 2024-12-13.

[2] National Cancer Institute. What is cancer?, 2021. Accessed: 2024-12-13.

[3] Eduardo Rosenblatt, Eduardo Zubizarreta, et al. Radiotherapy in cancer care: facing the global challenge. International Atomic Energy Agency Vienna, 2017.

[4] Josep M Borras, Yolande Lievens, Peter Dunscombe, Mary Coffey, Julian Malicki, Julieta Corral, Chiara Gasparotto, Noemie Defourny, Michael Barton, Rob Verhoeven, et al. The optimal utilization proportion of external beam radiotherapy in european countries: an estro-hero analysis. Radiotherapy and Oncology, 116(1):38–44, 2015.

[5] Geoff Delaney, Susannah Jacob, Carolyn Featherstone, and Michael Barton. The role of radiotherapy in cancer treatment: estimating optimal utilization from a review of evidence-based clinical guidelines. Cancer: Interdisciplinary International Journal of the American Cancer Society, 104(6):1129–1137, 2005.

[6] Rany Nuraini and Rena Widita. Tumor control probability (tcp) and normal tissue complication probability (ntcp) with consideration of cell biological effect. In Journal of Physics: Conference Series, page 012092. IOP Publishing, 2019.

[7] Sarah Baatout. Radiobiology textbook. Springer, 2023.

[8] SØREN M Bentzen. Dose-response relationships in radiotherapy. Basic clinical radiobiology, 4, 2002.

[9] David Thwaites. Accuracy required and achievable in radiotherapy dosimetry: have modern technology and techniques changed our views? In Journal of Physics: Conference Series, volume 444, page 012006. IOP Publishing, 2013.

[10] Douglas Jones. Icru report 50—prescribing, recording and reporting photon beam therapy, 1994.

[11] Ervin B Podgorsak et al. Review of radiation oncology physics: a handbook for teachers and students. Vienna, Austria: IAE Agency, 19:133, 2003.

[12] Jenny Bertholet, Gail Anastasi, David Noble, Arjan Bel, Ruud van Leeuwen, Toon Roggen, Michael Duchateau, Sara Pilskog, Cristina Garibaldi, Nina Tilly, et al. Patterns of practice for adaptive and real-time radiation therapy (pop-art rt) part ii: Offline and online plan adaption for interfractional changes. Radiotherapy and Oncology, 153:88–96, 2020.

[13] Thomas Bortfeld, Arthur L Boyer, Wolfgang Schlegel, Darren L Kahler, and Timothy J Waldron. Realization and verification of three-dimensional conformal radiotherapy with modulated fields. International Journal of Radiation Oncology* Biology* Physics, 30(4):899–908, 1994.

[14] Karl Otto. Volumetric modulated arc therapy: Imrt in a single gantry arc. Medical physics, 35(1):310–317, 2008.

[15] Vincent Grégoire, Matthias Guckenberger, Karin Haustermans, Jan JW Lagendijk, Cynthia Ménard, Richard Pötter, Ben J Slotman, Kari Tanderup, Daniela Thorwarth, Marcel Van Herk, et al. Image guidance in radiation therapy for better cure of cancer. Molecular oncology, 14(7):1470–1491, 2020.

[16] Christopher Kurz, Giulia Buizza, Guillaume Landry, Florian Kamp, Moritz Rabe, Chiara Paganelli, Guido Baroni, Michael Reiner, Paul J Keall, Cornelis AT van den Berg, et al. Medical physics challenges in clinical mr-guided radiotherapy. Radiation Oncology, 15:1–16, 2020.

[17] Jenny Bertholet, Antje Knopf, Björn Eiben, Jamie McClelland, Alexander Grimwood, Emma Harris, Martin Menten, Per Poulsen, Doan Trang Nguyen, Paul Keall, et al. Real-time intrafraction motion monitoring in external beam radiotherapy. Physics in medicine & biology, 64(15):15TR01, 2019.

[18] Jean-Paul JE Kleijnen, Bram Van Asselen, Johannes PM Burbach, Martijn Intven, Marielle EP Philippens, Onne Reerink, Jan JW Lagendijk, and Bas W Raaymakers. Evolution of motion uncertainty in rectal cancer: implications for adaptive radiotherapy. Physics in Medicine & Biology, 61(1):1, 2015.

[19] Cornel Zachiu, Baudouin Denis de Senneville, Chrit Moonen, and Mario Ries. A framework for the correction of slow physiological drifts during mr-guided hifu therapies: proof of concept. Medical physics, 42(7):4137–4148, 2015.

[20] Paul J Keall, Gig S Mageras, James M Balter, Richard S Emery, Kenneth M Forster, Steve B Jiang, Jeffrey M Kapatoes, Daniel A Low, Martin J Murphy, Brad R Murray, et al. The management of respiratory motion in radiation oncology report of aapm task group 76 a. Medical physics, 33(10):3874–3900, 2006.

[21] Gail Anastasi, Jenny Bertholet, Per Poulsen, Toon Roggen, Cristina Garibaldi, Nina Tilly, Jeremy T Booth, Uwe Oelfke, Ben Heijmen, and Marianne C Aznar. Patterns of practice for adaptive and real-time radiation therapy (pop-art rt) part i: Intra-fraction breathing motion management. Radiotherapy and Oncology, 153:79–87, 2020.

[22] Sonja Dieterich, Kevin Cleary, Warren D'Souza, Martin Murphy, Kenneth H Wong, and Paul Keall. Locating and targeting moving tumors with radiation beams. Medical physics, 35(12):5684–5694, 2008.

[23] Joep C Stroom and Ben JM Heijmen. Geometrical uncertainties, radiotherapy planning margins, and the icru-62 report. Radiotherapy and oncology, 64(1):75–83, 2002.

[24] Marcel Van Herk. Errors and margins in radiotherapy. In Seminars in radiation oncology, volume 14, pages 52–64. Elsevier, 2004.

[25] J Darréon, G Bouilhol, N Aillières, H Bouscayrol, L Simon, and M Ayadi. Respiratory motion management for external radiotherapy treatment. Cancer/Radiothérapie, 26(1-2):50–58, 2022.

[26] Jochem WH Wolthaus, Jan-Jakob Sonke, Marcel van Herk, José SA Belderbos, Maddalena MG Rossi, Joos V Lebesque, and Eugène MF Damen. Comparison of different strategies to use four-dimensional computed tomography in treatment planning for lung cancer patients. International Journal of Radiation Oncology* Biology* Physics, 70(4):1229–1238, 2008.

[27] Emma Colvill, Jeremy Booth, Simeon Nill, Martin Fast, James Bedford, Uwe Oelfke, Mitsuhiro Nakamura, Per Poulsen, Esben Worm, Rune Hansen, et al. A dosimetric comparison of real-time adaptive and non-adaptive radiotherapy: a multi-institutional study encompassing robotic, gimbaled, multileaf collimator and couch tracking. Radiotherapy and Oncology, 119(1):159–165, 2016.

[28] Cornelis Ph Kamerling, Martin F Fast, Peter Ziegenhein, Martin J Menten, Simeon Nill, and Uwe Oelfke. Real-time 4d dose reconstruction for tracked dynamic mlc deliveries for lung sbrt. Medical physics, 43(11):6072–6081, 2016.

[29] Saber Nankali, Esben S Worm, Rune Hansen, Britta Weber, Morten Høyer, Alireza Zirak, and Per Rugaard Poulsen. Geometric and dosimetric comparison of four intrafraction motion adaptation strategies for stereotactic liver radiotherapy. Physics in Medicine & Biology, 63(14):145010, 2018.

[30] Sebastian Klüter. Technical design and concept of a 0.35 t mr-linac. Clinical and translational radiation oncology, 18:98–101, 2019.

[31] Paul J Keall, Amit Sawant, Ross I Berbeco, Jeremy T Booth, Byungchul Cho, Laura I Cerviño, Eileen Cirino, Sonja Dieterich, Martin F Fast, Peter B Greer, et al. Aapm task group 264: The safe clinical implementation of mlc tracking in radiotherapy. Medical physics, 48(5):e44–e64, 2021.

[32] Markus Glitzner, PL Woodhead, PTS Borman, JJW Lagendijk, and BW Raaymakers. Mlc-tracking performance on the elekta unity mri-linac. Physics in Medicine & Biology, 64(15):15NT02, 2019.

[33] Paul ZY Liu, Bing Dong, Doan Trang Nguyen, Yuanyuan Ge, Emily A Hewson, David EJ Waddington, Ricky O'Brien, Gary P Liney, and Paul J Keall. First experimental investigation of simultaneously tracking two independently moving targets on an mri-linac using real-time mri and mlc tracking. Medical Physics, 47(12):6440–6449, 2020.

[34] Tom Depuydt, Kenneth Poels, Dirk Verellen, Benedikt Engels, Christine Collen, Manuela Buleteanu, Robbe Van den Begin, Marlies Boussaer, Michael Duchateau, Thierry Gevaert, et al. Treating patients with real-time tumor tracking using the vero gimbaled linac system: implementation and first review. Radiotherapy and Oncology, 112(3):343–351, 2014.

[35] Esben S Worm, Morten Høyer, Rune Hansen, Lars P Larsen, Britta Weber, Cai Grau, and Per R Poulsen. A prospective cohort study of gated stereotactic liver radiation therapy using continuous internal electromagnetic motion monitoring. International Journal of Radiation Oncology* Biology* Physics, 101(2):366–375, 2018.

[36] Stefanie Ehrbar, Rosalind Perrin, Marta Peroni, Kinga Bernatowicz, Thomas Parkel, Izabela Pytko, Stephan Klöck, Matthias Guckenberger, Stephanie Tanadini-Lang, Damien Charles Weber, et al. Respiratory motion-management in stereotactic body radiation therapy for lung cancer–a dosimetric comparison in an anthropomorphic lung phantom (luca). Radiotherapy and oncology, 121(2):328–334, 2016.

[37] Mischa Hoogeman, Jean-Briac Prévost, Joost Nuyttens, Johan Pöll, Peter Levendag, and Ben Heijmen. Clinical accuracy of the respiratory tumor tracking system of the cyberknife: assessment by analysis of log files. International Journal of Radiation Oncology* Biology* Physics, 74(1):297–303, 2009.

[38] Brian K Chang and Robert D Timmerman. Stereotactic body radiation therapy: a comprehensive review. American journal of clinical oncology, 30(6):637–644, 2007.

[39] Markus Glitzner, SPM Crijns, B Denis De Senneville, Charis Kontaxis, FM Prins, JJW Lagendijk, and BW Raaymakers. On-line mr imaging for dose validation of abdominal radiotherapy. Physics in Medicine & Biology, 60(22):8869, 2015.

[40] Bjorn Stemkens, Markus Glitzner, Charis Kontaxis, Baudouin Denis De Senneville, Fieke M Prins, Sjoerd PM Crijns, Linda GW Kerkmeijer, Jan JW Lagendijk, Cornelis AT Van Den Berg, and Rob HN Tijssen. Effect of intra-fraction motion on the accumulated dose for free-breathing mr-guided stereotactic body radiation therapy of renal-cell carcinoma. Physics in Medicine & Biology, 62(18):7407, 2017.

[41] Charis Kontaxis, GH Bol, JJW Lagendijk, and BW Raaymakers. A new methodology for inter-and intrafraction plan adaptation for the mr-linac. Physics in Medicine & Biology, 60(19):7485, 2015.

[42] PTS Borman, C Bos, B Stemkens, CTW Moonen, BW Raaymakers, and RHN Tijssen. Assessment of 3d motion modeling performance for dose accumulation mapping on the mr-linac by simultaneous multislice mri. Physics in Medicine & Biology, 64(9):095004, 2019.

[43] Jian-Yue Jin, Fang-Fang Yin, Stephen E Tenn, Paul M Medin, and Timothy D Solberg. Use of the brainlab exactrac x-ray 6d system in image-guided radiotherapy. Medical Dosimetry, 33(2):124–134, 2008.

[44] Karl Rasmussen, Victoria Bry, and Nikos Papanikolaou. Technical overview and features of the c-rad catalyst™ and sentinel™ systems. Surface Guided Radiation Therapy, pages 51–72, 2020.

[45] M Ducassou, D Marre, N Mathy, J Mazurier, P Navarro, D Zarate, C Chevelle, P Dudouet, D Franck, O Gallocher, et al. Quality control of the imaging and repositioning system of vero accelerator (brainlab-mitsubishi) for stereotactic treatments. Physica Medica: European Journal of Medical Physics, 31:e38–e39, 2015.

[46] William C Welch and Peter C Gerszten. Accuray cyberknife® image-guided radiosurgical system. Expert Review of Medical Devices, 2(2):141–147, 2005.

[47] Olga L Green, Leith J Rankine, Bin Cai, Austen Curcuru, Rojano Kashani, Vivian Rodriguez, H Harold Li, Parag J Parikh, Clifford G Robinson, Jeffrey R Olsen, et al. First clinical implementation of real-time, real anatomy tracking and radiation beam control. Medical physics, 45(8):3728–3740, 2018.

[48] Guillaume Landry and Chia-ho Hua. Current state and future applications of radiological image guidance for particle therapy. Medical Physics, 45(11):e1086–e1095, 2018.

[49] Dirk Verellen, Mark De Ridder, and Guy Storme. A (short) history of image-guided radiotherapy. Radiotherapy and Oncology, 86(1):4–13, 2008.

[50] Kavitha Srinivasan, Mohammad Mohammadi, and Justin Shepherd. Applications of linac-mounted kilovoltage cone-beam computed tomography in modern radiation therapy: A review. Polish journal of radiology, 79:181, 2014.

[51] Ajay R Bhoosreddy and Priyanka Umesh Sakhavalkar. Image deteriorating factors in cone beam computed tomography, their classification, and measures to reduce them: A pictorial essay. Journal of Indian Academy of Oral Medicine and Radiology, 26(3):293–297, 2014.

[52] Parham Alaei and Emiliano Spezi. Imaging dose from cone beam computed tomography in radiation therapy. Physica Medica, 31(7):647–658, 2015.

[53] Elisabeth Steiner, Chun-Chien Shieh, Vincent Caillet, Jeremy Booth, Ricky O'Brien, Adam Briggs, Nicholas Hardcastle, Dasantha Jayamanne, Kathryn Szymura, Thomas Eade, et al. Both four-dimensional computed tomography and four-dimensional cone beam computed tomography under-predict lung target motion during radiotherapy. Radiotherapy and Oncology, 135:65–73, 2019.

[54] Marcel van Herk, Alan McWilliam, Michael Dubec, Corinne Faivre-Finn, and Ananya Choudhury. Magnetic resonance imaging–guided radiation therapy: a short strengths, weaknesses, opportunities, and threats analysis, 2018.

[55] David Anthony Jaffray, Caroline Chung, Catherine Coolens, Warren Foltz, Harald Keller, Cynthia Menard, Michael Milosevic, Julia Publicover, and Ivan Yeung. Quantitative imaging in radiation oncology: an emerging science and clinical service. In Seminars in radiation oncology, volume 25, pages 292–304. Elsevier, 2015.

[56] Ayshea Hameeduddin and Anju Sahdev. Diffusion-weighted imaging and dynamic contrast-enhanced mri in assessing response and recurrent disease in gynaecological malignancies. Cancer Imaging, 15:1–12, 2015.

[57] Martin J Menten, Andreas Wetscherek, and Martin F Fast. Mri-guided lung sbrt: present and future developments. Physica Medica, 44:139–149, 2017.

[58] Hee Jung Shin, Hak Hee Kim, Ki Chang Shin, Yoo Sub Sung, Joo Hee Cha, Jong Won Lee, Byung Ho Son, and Sei Hyun Ahn. Prediction of low-risk breast

cancer using perfusion parameters and apparent diffusion coefficient. Magnetic Resonance Imaging, 34(2):67–74, 2016.

[59] Magda Ali Hany El Bakry, Amina Ahmed Sultan, Nahed Abd Elgaber El-Tokhy, Tamer Fady Yossif, and Carmen Ali Ahmed Ali. Role of diffusion weighted imaging and dynamic contrast enhanced magnetic resonance imaging in breast tumors. The Egyptian Journal of Radiology and Nuclear Medicine, 46(3):791–804, 2015.

[60] Hans-Ulrich Kauczor, Christian Zechmann, Bram Stieltjes, and Marc-Andre Weber. Functional magnetic resonance imaging for defining the biological target volume. Cancer Imaging, 6(1):51, 2006.

[61] Uulke A Van der Heide, Antonetta C Houweling, Greetje Groenendaal, Regina GH Beets-Tan, and Philippe Lambin. Functional mri for radiotherapy dose painting. Magnetic resonance imaging, 30(9):1216–1223, 2012.

[62] Harriet C Thoeny and Brian D Ross. Predicting and monitoring cancer treatment response with diffusion-weighted mri. Journal of Magnetic Resonance Imaging, 32(1):2–16, 2010.

[63] CJ Galbán, BA Hoff, TL Chenevert, and BD Ross. Diffusion mri in early cancer therapeutic response assessment. NMR in biomedicine, 30(3):e3458, 2017.

[64] Daniela Thorwarth, Mike Notohamiprodjo, Daniel Zips, and Arndt-Christan Müller. Personalized precision radiotherapy by integration of multi-parametric functional and biological imaging in prostate cancer: a feasibility study. Zeitschrift für Medizinische Physik, 27(1):21–30, 2017.

[65] Sasa Mutic and James F Dempsey. The viewray system: magnetic resonance–guided and controlled radiotherapy. In Seminars in radiation oncology, volume 24, pages 196–199. Elsevier, 2014.

[66] Jan JW Lagendijk, Bas W Raaymakers, and Marco Van Vulpen. The magnetic resonance imaging–linac system. In Seminars in radiation oncology, volume 24, pages 207–209. Elsevier, 2014.

[67] B Gino Fallone, Satyapal Rathee, Nicola de Zanche, Eugene Yip, Keith Wachowicz, and Jihyun Yun. The alberta rotating biplanar linac-mr, aka, aurora-rt™. In A Practical Guide to MR-Linac: Technical Innovation and Clinical Implication, pages 193–215. Springer, 2024.

[68] Dennis Winkel, Gijsbert H Bol, Petra S Kroon, Bram van Asselen, Sara S Hackett, Anita M Werensteijn-Honingh, Martijn PW Intven, Wietse SC Eppinga, Rob HN

Tijssen, Linda GW Kerkmeijer, et al. Adaptive radiotherapy: the elekta unity mr-linac concept. Clinical and translational radiation oncology, 18:54–59, 2019.

[69] Paul J Keall, Michael Barton, Stuart Crozier, et al. The australian magnetic resonance imaging–linac program. In Seminars in radiation oncology, volume 24, pages 203–206. Elsevier, 2014.

[70] Gary P Liney, B Whelan, B Oborn, Michael Barton, and P Keall. Mri-linear accelerator radiotherapy systems. Clinical Oncology, 30(11):686–691, 2018.

[71] Inc. ViewRay Systems. Viewray systems. https://viewraysystems.com/, 2024. Accessed: 2024-12-27.

[72] Elekta. Unity. https://www.elekta.com/products/radiation-therapy/unity/, 2024. Accessed: 2024-12-27.

[73] MagnetTx Oncology Solutions Ltd. Aurora-rt. https://www.magnettx.com/aurora-rt, 2024. Accessed: 2024-12-27.

[74] Jeffrey Olsen, Olga Green, and Rojano Kashani. World's first applicaton of mr-guidance for radiotherapy. Missouri medicine, 112(5):358, 2015.

[75] Bas W Raaymakers, IM Jürgenliemk-Schulz, GH Bol, M Glitzner, ANTJ Kotte, B Van Asselen, JCJ De Boer, JJ Bluemink, SL Hackett, MA Moerland, et al. First patients treated with a 1.5 t mri-linac: clinical proof of concept of a high-precision, high-field mri guided radiotherapy treatment. Physics in Medicine & Biology, 62(23):L41, 2017.

[76] Oliver Bieri and Klaus Scheffler. Fundamentals of balanced steady state free precession mri. Journal of Magnetic Resonance Imaging, 38(1):2–11, 2013.

[77] Alexandra E Bourque, Stéphane Bedwani, Jean-François Carrier, Cynthia Ménard, Pim Borman, Clemens Bos, Bas W Raaymakers, Nikolai Mickevicius, Eric Paulson, and Rob HN Tijssen. Particle filter–based target tracking algorithm for magnetic resonance–guided respiratory compensation: robustness and accuracy assessment. International Journal of Radiation Oncology* Biology* Physics, 100(2):325–334, 2018.

[78] Bjorn Stemkens, Eric S Paulson, and Rob HN Tijssen. Nuts and bolts of 4d-mri for radiotherapy. Physics in Medicine & Biology, 63(21):21TR01, 2018.

[79] Chiara Paganelli, B Whelan, Marta Peroni, Paul Summers, M Fast, Tessa van de Lindt, J McClelland, Björn Eiben, P Keall, T Lomax, et al. Mri-guidance for motion management in external beam radiotherapy: current status and future challenges. Physics in Medicine & Biology, 63(22):22TR03, 2018.

[80] Taeho Kim, Benjamin Lewis, Rajiv Lotey, Enzo Barberi, and Olga Green. Clinical experience of mri4d quasar motion phantom for latency measurements in 0.35 t mr-linac. Journal of applied clinical medical physics, 22(1):128–136, 2021.

[81] Andrzej P Wojcieszynski, Stephen A Rosenberg, Jeffrey V Brower, Craig R Hullett, Mark W Geurts, Zacariah E Labby, Patrick M Hill, R Adam Bayliss, Bhudatt Paliwal, John E Bayouth, et al. Gadoxetate for direct tumor therapy and tracking with real-time mri-guided stereotactic body radiation therapy of the liver. Radiotherapy and Oncology, 118(2):416–418, 2016.

[82] Lauren E Henke, JA Contreras, OL Green, B Cai, H Kim, MC Roach, JR Olsen, B Fischer-Valuck, DF Mullen, R Kashani, et al. Magnetic resonance image-guided radiotherapy (mrigrt): a 4.5-year clinical experience. Clinical Oncology, 30(11):720–727, 2018.

[83] John R van Sörnsen de Koste, Miguel A Palacios, Anna ME Bruynzeel, Ben J Slotman, Suresh Senan, and Frank J Lagerwaard. Mr-guided gated stereotactic radiation therapy delivery for lung, adrenal, and pancreatic tumors: a geometric analysis. International Journal of Radiation Oncology* Biology* Physics, 102(4):858–866, 2018.

[84] Tobias Finazzi, Miguel A Palacios, Cornelis JA Haasbeek, Marjan A Admiraal, Femke OB Spoelstra, Anna ME Bruynzeel, Berend J Slotman, Frank J Lagerwaard, and Suresh Senan. Stereotactic mr-guided adaptive radiation therapy for peripheral lung tumors. Radiotherapy and Oncology, 144:46–52, 2020.

[85] Andreas Wedel, T Pock, C Zach, H Bischof, and D Cremers. An improved algorithm for tv-l 1 optical flow. statistical and geometrical approaches to visual motion analysis, 23-45, 2009.

[86] Panpan Hu, Xiaoyang Li, Wei Liu, Bing Yan, Xudong Xue, Fei Yang, John Chetley Ford, Lorraine Portelance, and Yidong Yang. Dosimetry impact of gating latency in cine magnetic resonance image guided breath-hold pancreatic cancer radiotherapy. Physics in Medicine & Biology, 67(5):055008, 2022.

[87] James L Bedford, Martin F Fast, Simeon Nill, Fiona MA McDonald, Merina Ahmed, Vibeke N Hansen, and Uwe Oelfke. Effect of mlc tracking latency on conformal volumetric modulated arc therapy (vmat) plans in 4d stereotactic lung treatment. Radiotherapy and Oncology, 117(3):491–495, 2015.

[88] PTS Borman, RHN Tijssen, C Bos, CTW Moonen, BW Raaymakers, and M Glitzner. Characterization of imaging latency for real-time mri-guided radiotherapy. Physics in Medicine & Biology, 63(15):155023, 2018.

[89] Sean CL Deoni, Brian K Rutt, and Terry M Peters. Rapid combined t1 and t2 mapping using gradient recalled acquisition in the steady state. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 49(3):515–526, 2003.

[90] Zizhao Zhang, Adriana Romero, Matthew J Muckley, Pascal Vincent, Lin Yang, and Michal Drozdzal. Reducing uncertainty in undersampled mri reconstruction with active acquisition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2049–2058, 2019.

[91] Seul Lee, Soozy Jung, Kyu-Jin Jung, and Dong-Hyun Kim. Deep learning in mr motion correction: a brief review and a new motion simulation tool (view2dmotion). Investigative Magnetic Resonance Imaging, 24(4):196–206, 2020.

[92] Frank Godenschweger, Urte Kägebein, Daniel Stucht, Uten Yarach, Alessandro Sciarra, Renat Yakupov, Falk Lüsebrink, Peter Schulze, and Oliver Speck. Motion correction in mri of the brain. Physics in medicine & biology, 61(5):R32, 2016.

[93] Vahid Ghodrati, Mark Bydder, Fadil Ali, Chang Gao, Ashley Prosper, Kim-Lien Nguyen, and Peng Hu. Retrospective respiratory motion correction in cardiac cine mri reconstruction using adversarial autoencoder and unsupervised learning. NMR in Biomedicine, 34(2):e4433, 2021.

[94] Patricia M Johnson and Maria Drangova. Conditional generative adversarial network for 3d rigid-body motion correction in mri. Magnetic resonance in medicine, 82(3):901–910, 2019.

[95] Philippe Giraud, Sabine Elles, Sylvie Helfre, Yann De Rycke, Vincent Servois, Marie-France Carette, Claude Alzieu, Pierre-Yves Bondiau, Bernard Dubray, Emmanuel Touboul, et al. Conformal radiotherapy for lung cancer: different delineation of the gross tumor volume (gtv) by radiologists and radiation oncologists. Radiotherapy and oncology, 62(1):27–36, 2002.

[96] Ting Chen, Songbing Qin, Xiaoting Xu, Salma K Jabbour, Bruce G Haffty, and Ning J Yue. Frequency filtering based analysis on the cardiac induced lung tumor motion and its impact on the radiotherapy management. Radiotherapy and Oncology, 112(3):365–370, 2014.

[97] Mai Lykkegaard Schmidt, Lone Hoffmann, Marianne Marquard Knap, Torben Riis Rasmussen, Birgitte Holst Folkersen, Jakob Toftegaard, Ditte Sloth Møller, and Per Rugård Poulsen. Cardiac and respiration induced motion of mediastinal

lymph node targets in lung cancer patients throughout the radiotherapy treatment course. Radiotherapy and Oncology, 121(1):52–58, 2016.

[98] Yvette Seppenwoolde, Hiroki Shirato, Kei Kitamura, Shinichi Shimizu, Marcel Van Herk, Joos V Lebesque, and Kazuo Miyasaka. Precise and real-time measurement of 3d tumor motion in lung due to breathing and heartbeat, measured during radiotherapy. International Journal of Radiation Oncology* Biology* Physics, 53(4):822–834, 2002.

[99] Hiroki Shirato, Keishiro Suzuki, Gregory C Sharp, Katsuhisa Fujita, Rikiya Onimaru, Masaharu Fujino, Norio Kato, Yasuhiro Osaka, Rumiko Kinoshita, Hiroshi Taguchi, et al. Speed and amplitude of lung tumor motion precisely detected in four-dimensional setup and in real-time tumor-tracking radiotherapy. International Journal of Radiation Oncology* Biology* Physics, 64(4):1229–1236, 2006.

[100] Alan McWilliam, Jason Kennedy, Clare Hodgson, Eliana Vasquez Osorio, Corinne Faivre-Finn, and Marcel Van Herk. Radiation dose to heart base linked with poorer survival in lung cancer patients. European Journal of Cancer, 85:106–113, 2017.

[101] Elia Lombardo, Moritz Rabe, Yuqing Xiong, Lukas Nierer, Davide Cusumano, Lorenzo Placidi, Luca Boldrini, Stefanie Corradini, Maximilian Niyazi, Claus Belka, et al. Offline and online lstm networks for respiratory motion prediction in mr-guided radiotherapy. Physics in Medicine & Biology, 67(9):095006, 2022.

[102] Jihyun Yun, Marc Mackenzie, Satyapal Rathee, Don Robinson, and BG Fallone. An artificial neural network (ann)-based lung-tumor motion predictor for intrafractional mr tumor tracking. Medical physics, 39(7Part1):4423–4433, 2012.

[103] Andreas Krauss, Simeon Nill, and U Oelfke. The comparative performance of four respiratory motion predictors for real-time tumour tracking. Physics in Medicine & Biology, 56(16):5303, 2011.

[104] Elia Lombardo, Moritz Rabe, Yuqing Xiong, Lukas Nierer, Davide Cusumano, Lorenzo Placidi, Luca Boldrini, Stefanie Corradini, Maximilian Niyazi, Michael Reiner, et al. Evaluation of real-time tumor contour prediction using lstm networks for mr-guided radiotherapy. Radiotherapy and Oncology, 182:109555, 2023.

[105] Lianlei Shan, Xiaobin Li, and Weiqiang Wang. Decouple the high-frequency and low-frequency information of images for semantic segmentation. In ICASSP

2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1805–1809. IEEE, 2021.

[106] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16, pages 435–452. Springer, 2020.

[107] Shu Tang, Haiheng Ran, Shuli Yang, Zhaoxia Wang, Wei Li, Haorong Li, and Zihao Meng. A frequency selection network for medical image segmentation. Heliyon, 10(16), 2024.

[108] Chang Gao, Shu-Fu Shih, J Paul Finn, and Xiaodong Zhong. A projection-based k-space transformer network for undersampled radial mri reconstruction with limited training subjects. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 726–736. Springer, 2022.

[109] Maarten L Terpstra, Matteo Maspero, Federico d'Agata, Bjorn Stemkens, Martijn PW Intven, Jan JW Lagendijk, Cornelis AT Van den Berg, and Rob HN Tijssen. Deep learning-based image reconstruction and motion estimation from undersampled radial k-space for real-time mri-guided radiotherapy. Physics in Medicine & Biology, 65(15):155015, 2020.

[110] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pages 694–711. Springer, 2016.

[111] SN Friedman and IA Cunningham. A moving slanted-edge method to measure the temporal modulation transfer function of fluoroscopic systems. Medical physics, 35(6Part1):2473–2484, 2008.

[112] JA Rowlands. Videofluorography: The role of temporal averaging. Medical physics, 11(2):129–136, 1984.

[113] Yabo Fu, Thomas R Mazur, Xue Wu, Shi Liu, Xiao Chang, Yonggang Lu, H Harold Li, Hyun Kim, Michael C Roach, Lauren Henke, et al. A novel mri segmentation method using cnn-based correction network for mri-guided adaptive radiotherapy. Medical physics, 45(11):5129–5137, 2018.

[114] Maria Kawula, Indrawati Hadi, Lukas Nierer, Marica Vagni, Davide Cusumano, Luca Boldrini, Lorenzo Placidi, Stefanie Corradini, Claus Belka, Guillaume

Landry, et al. Patient-specific transfer learning for auto-segmentation in adaptive 0.35 t mrgrt of prostate cancer: a bi-centric evaluation. Medical Physics, 50(3):1573–1585, 2023.

[115] Hajar Emami, Ming Dong, Siamak P Nejad-Davarani, and Carri K Glide-Hurst. Generating synthetic cts from magnetic resonance images using generative adversarial networks. Medical physics, 45(8):3627–3636, 2018.

[116] Chenyang Shen, Dan Nguyen, Liyuan Chen, Yesenia Gonzalez, Rafe McBeth, Nan Qin, Steve B Jiang, and Xun Jia. Operating a treatment planning system using a deep-reinforcement learning-based virtual treatment planner for prostate cancer intensity-modulated radiation therapy treatment planning. Medical physics, 47(6):2329–2336, 2020.

[117] Davide Cusumano, Luca Boldrini, Jennifer Dhont, Claudio Fiorino, Olga Green, Görkem Güngör, Núria Jornet, Sebastian Klüter, Guillaume Landry, Gian Carlo Mattiucci, et al. Artificial intelligence in magnetic resonance guided radiotherapy: Medical and physical considerations on state of art and future perspectives. Physica medica, 85:175–191, 2021.

[118] Robert W Brown, Y-C Norman Cheng, E Mark Haacke, Michael R Thompson, and Ramesh Venkatesan. Magnetic resonance imaging: physical principles and sequence design. John Wiley & Sons, 2014.

[119] GM Bydder, JV Hajnal, and IR Young. Mri: use of the inversion recovery pulse sequence. Clinical radiology, 53(3):159–176, 1998.

[120] Jiayu Song, Yanhui Liu, Sally L Gewalt, Gary Cofer, G Allan Johnson, and Qing Huo Liu. Least-square nufft methods applied to 2-d and 3-d radially encoded mr image reconstruction. IEEE Transactions on Biomedical Engineering, 56(4):1134–1142, 2009.

[121] Liewei Sha, Hua Guo, and Allen W Song. An improved gridding method for spiral mri using nonuniform fast fourier transform. Journal of Magnetic Resonance, 162(2):250–258, 2003.

[122] John M Pauly. Gridding & the nufft for non-cartesian image reconstruction. ISMRM Educational Course on Image Reconstruction, 45, 2012.

[123] John I Jackson, Craig H Meyer, Dwight G Nishimura, and Albert Macovski. Selection of a convolution function for fourier inversion using gridding (computerised tomography application). IEEE transactions on medical imaging, 10(3):473–478, 1991.

[124] Camila Munoz, Anastasia Fotaki, René M Botnar, and Claudia Prieto. Latest advances in image acceleration: all dimensions are fair game. Journal of Magnetic Resonance Imaging, 57(2):387–402, 2023.

[125] G McGibney, MR Smith, ST Nichols, and A Crawley. Quantitative evaluation of several partial fourier reconstruction algorithms used in mri. Magnetic resonance in medicine, 30(1):51–59, 1993.

[126] Klaas P Pruessmann, Markus Weiger, Markus B Scheidegger, and Peter Boesiger. Sense: sensitivity encoding for fast mri. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 42(5):952–962, 1999.

[127] Mark A Griswold, Peter M Jakob, Robin M Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase. Generalized autocalibrating partially parallel acquisitions (grappa). Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 47(6):1202–1210, 2002.

[128] David L Donoho. Compressed sensing. IEEE Transactions on information theory, 52(4):1289–1306, 2006.

[129] Alice C Yang, Madison Kretzler, Sonja Sudarski, Vikas Gulani, and Nicole Seiberlich. Sparse reconstruction techniques in magnetic resonance imaging. Investigative Radiology, 51(6):349–364, 2016.

[130] Chang Min Hyun, Hwa Pyung Kim, Sung Min Lee, Sungchul Lee, and Jin Keun Seo. Deep learning for undersampled mri reconstruction. Physics in Medicine & Biology, 63(13):135007, 2018.

[131] Ziheng Zhao, Tianjiao Zhang, Weidi Xie, Yanfeng Wang, and Ya Zhang. K-space transformer for undersampled mri reconstruction. arXiv preprint arXiv:2206.06947, 2022.

[132] Alexander Wu Chao, Maury Tigner, Hans Weise, and Frank Zimmermann. Handbook of accelerator physics and engineering. World scientific, 2023.

[133] Elekta. Elekta - precision radiation medicine, 2025. Accessed: 2025-03-06.

[134] Wolfgang Schlegel, Christian P Karger, and Oliver Jäkel. Medizinische Physik: Grundlagen–Bildgebung–Therapie–Technik. Springer-Verlag, 2018.

[135] Uulke van der Heide and David I Thwaites. Integrated mri-linac systems: The new paradigm for precision adaptive radiotherapy and biological image-guidance? Radiotherapy and Oncology, 176:249–250, 2022.

[136] S Klüter. Technical design and concept of a 0.35 t mr-linac. clin transl radiat oncol 2019; 18: 98–101.

[137] Zhuojie Sui, Prasannakumar Palaniappan, Jakob Brenner, Chiara Paganelli, Christopher Kurz, Guillaume Landry, and Marco Riboldi. Intra-frame motion deterioration effects and deep-learning-based compensation in mr-guided radiotherapy. Medical Physics, 51(3):1899–1917, 2024.

[138] Mark A Griswold, Peter M Jakob, Mathias Nittka, James W Goldfarb, and Axel Haase. Partially parallel imaging with localized sensitivities (pils). Magnetic resonance in medicine, 44(4):602–609, 2000.

[139] Li Feng. Golden-angle radial mri: basics, advances, and applications. Journal of Magnetic Resonance Imaging, 56(1):45–62, 2022.

[140] Stefan Wundrak, Jan Paul, Johannes Ulrici, Erich Hell, and Volker Rasche. A small surrogate for the golden angle in time-resolved radial mri based on generalized fibonacci sequences. IEEE transactions on medical imaging, 34(6):1262–1269, 2014.

[141] Stefan Wundrak, Jan Paul, Johannes Ulrici, Erich Hell, Margrit-Ann Geibel, Peter Bernhardt, Wolfgang Rottbauer, and Volker Rasche. Golden ratio sparse mri using tiny golden angles. Magnetic resonance in medicine, 75(6):2372–2378, 2016.

[142] Gastao Cruz, Kerstin Hammernik, Thomas Kuestner, Carlos Velasco, Alina Hua, Tevfik Fehmi Ismail, Daniel Rueckert, Rene Michael Botnar, and Claudia Prieto. Single-heartbeat cardiac cine imaging via jointly regularized nonrigid motion-corrected reconstruction. NMR in Biomedicine, 36(9):e4942, 2023.

[143] Stephen J Riederer, Talin Tasciyan, Farhad Farzaneh, James N Lee, Ronald C Wright, and Robert J Herfkens. Mr fluoroscopy: technical feasibility. Magnetic resonance in medicine, 8(1):1–15, 1988.

[144] Hee Kwon Song and Lawrence Dougherty. Dynamic mri with projection reconstruction and kwic processing for simultaneous high spatial and temporal resolution. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 52(4):815–824, 2004.

[145] Stefanie Winkelmann, Tobias Schaeffter, Thomas Koehler, Holger Eggers, and Olaf Doessel. An optimal radial profile order based on the golden ratio for time-resolved mri. IEEE transactions on medical imaging, 26(1):68–76, 2006.

[146] U. Computerized Imaging Reference Systems Inc. Cirs dynamicthorax phantom model 008a, 2024. Available online: https://www.cirsinc.com/.

[147] Moritz Rabe, Chiara Paganelli, Marco Riboldi, David Bondesson, Moritz Jörg Schneider, Thomas Chmielewski, Guido Baroni, Julien Dinkel, Michael Reiner, Guillaume Landry, et al. Porcine lung phantom-based validation of estimated 4d-mri using orthogonal cine imaging for low-field mr-linacs. Physics in Medicine & Biology, 66(5):055006, 2021.

[148] Dominik F Bauer, Tom Russ, Barbara I Waldkirch, Christian Tönnes, William P Segars, Lothar R Schad, Frank G Zöllner, and Alena-Kathrin Golla. Generation of annotated multimodal ground truth datasets for abdominal medical image registration. International journal of computer assisted radiology and surgery, 16:1277–1285, 2021.

[149] Chiara Paganelli, Paul Summers, Chiara Gianoli, Massimo Bellomi, Guido Baroni, and Marco Riboldi. A tool for validating mri-guided strategies: a digital breathing ct/mri phantom of the abdominal site. Medical & Biological Engineering & Computing, 55:2001–2014, 2017.

[150] W Paul Segars, G Sturgeon, S Mendonca, Jason Grimes, and Benjamin MW Tsui. 4d xcat phantom for multimodality imaging research. Medical physics, 37(9):4902–4915, 2010.

[151] Chiara Paganelli, S Portoso, N Garau, G Meschini, R Via, G Buizza, P Keall, M Riboldi, and G Baroni. Time-resolved volumetric mri in mri-guided radiotherapy: an in silico comparative analysis. Physics in Medicine & Biology, 64(18):185013, 2019.

[152] Thomas G Perkins and Felix W Wehrli. Csf signal enhancement in short tr gradient echo images. Magnetic resonance imaging, 4(6):465–467, 1986.

[153] Friedrich Fuchs, Gerhard Laub, and Kuni Othomo. Truefisp—technical considerations and cardiovascular applications. European journal of radiology, 46(1):28–32, 2003.

[154] SC Deoni, JA Kost, PA Adams, E O'Riordan, and BK Rutt. Quantification of liver iron with rapid 3d r1 and r2 mapping with despot1 and despot2. In Proc Int Soc Magn Reson Med, volume 11, page 889, 2004.

[155] Kim E Barrett, Susan M Barman, Scott Boitano, and Heddwen L Brooks. Ganong's review of medical physiology. McGraw-Hill Companies, Inc., 2010.

[156] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. Insights into imaging, 9:611–629, 2018.

[157] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 458–467, 2021.

[158] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8684–8694, 2020.

[159] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015.

[160] Harry Nyquist. Certain topics in telegraph transmission theory. Transactions of the American Institute of Electrical Engineers, 47(2):617–644, 1928.

[161] Orhun Utku Aydin, Abdel Aziz Taha, Adam Hilbert, Ahmed A Khalil, Ivana Galinovic, Jochen B Fiebach, Dietmar Frey, and Vince Istvan Madai. On the usage of average hausdorff distance for segmentation performance assessment: hidden error when used for ranking. European radiology experimental, 5:1–7, 2021.

[162] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.

[163] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825, 2017.

[164] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 2019.

[165] P Kingma Diederik. Adam: A method for stochastic optimization. (No Title), 2014.

[166] Zhuojie Sui, Prasannakumar Palaniappan, Chiara Paganelli, Christopher Kurz, Guillaume Landry, and Marco Riboldi. Imaging error reduction in radial cine-mri with deep learning-based intra-frame motion compensation. Physics in Medicine & Biology, 69(22):225011, 2024.

[167] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need.(nips), 2017. arXiv preprint arXiv:1706.03762, 10:S0140525X16001837, 2017.

[168] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In International Conference on Machine Learning, pages 10524–10533. PMLR, 2020.

[169] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7890–7899, 2020.

[170] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, volume 2, pages 1398–1402. Ieee, 2003.

[171] Matthew J Muckley, Ruben Stern, Tullie Murrell, and Florian Knoll. Torchkbnufft: A high-level, hardware-agnostic non-uniform fast fourier transform. In ISMRM Workshop on Data Sampling & Image Reconstruction, volume 22, 2020.

[172] Ricardo Llugsi, Samira El Yacoubi, Allyx Fontaine, and Pablo Lupera. Comparison between adam, adamax and adam w optimizers to implement a weather forecast based on neural networks for the andean city of quito. In 2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM), pages 1–6. IEEE, 2021.

[173] I Loshchilov. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[174] Tom Blöcker, Elia Lombardo, Sebastian N Marschner, Claus Belka, Stefanie Corradini, Miguel A Palacios, Marco Riboldi, Christopher Kurz, and Guillaume Landry. Mrgrt real-time target localization using foundation models for contour point tracking and promptable mask refinement. Physics in Medicine & Biology, 70(1):015004, 2024.

[175] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7482–7491, 2018.

[176] Rick Groenendijk, Sezer Karaoglu, Theo Gevers, and Thomas Mensink. Multi-loss weighting with coefficient of variations. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 1469–1478, 2021.

[177] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6836–6846, 2021.

[178] Joseph Weygand, Clifton David Fuller, Geoffrey S Ibbott, Abdallah SR Mohamed, Yao Ding, Jinzhong Yang, Ken-Pin Hwang, and Jihong Wang. Spatial precision in magnetic resonance imaging–guided radiation therapy: the role of geometric distortion. International Journal of Radiation Oncology* Biology* Physics, 95(4):1304–1316, 2016.

[179] Melissa W Haskell, Jon-Fredrik Nielsen, and Douglas C Noll. Off-resonance artifact correction for mri: A review. NMR in Biomedicine, 36(5):e4867, 2023.

[180] Junghyun Roh, Dongmin Ryu, and Jimin Lee. Ct synthesis with deep learning for mr-only radiotherapy planning: a review. Biomedical Engineering Letters, 14(6):1259–1278, 2024.

[181] Xiao Han. Mr-based synthetic ct generation using a deep convolutional neural network method. Medical physics, 44(4):1408–1419, 2017.

[182] Matteo Seregni, Chiara Paganelli, D Lee, PB Greer, Guido Baroni, PJ Keall, and Marco Riboldi. Motion prediction in mri-guided radiotherapy based on interleaved orthogonal cine-mri. Physics in Medicine & Biology, 61(2):872, 2016.

[183] Ran Wang, Xiaokun Liang, Xuanyu Zhu, and Yaoqin Xie. A feasibility of respiration prediction based on deep bi-lstm for real-time tumor tracking. IEEE Access, 6:51262–51268, 2018.

[184] Aswin Hoffmann, Bradley Oborn, Maryam Moteabbed, Susu Yan, Thomas Bortfeld, Antje Knopf, Herman Fuchs, Dietmar Georg, Joao Seco, Maria Francesca Spadea, et al. Mr-guided proton therapy: a review and a preview. Radiation Oncology, 15:1–13, 2020.

[185] Martin von Siebenthal, Gabor Szekely, Urs Gamper, Peter Boesiger, Antony Lomax, and Ph Cattin. 4d mr imaging of respiratory organ motion and its variability. Physics in Medicine & Biology, 52(6):1547, 2007.

[186] Marco Durante, Roberto Orecchia, and Jay S Loeffler. Charged-particle therapy in cancer: clinical uses and future perspectives. Nature Reviews Clinical Oncology, 14(8):483–495, 2017.

[187] Radhe Mohan. A review of proton therapy–current status and future directions. Precision radiation oncology, 6(2):164–176, 2022.

[188] Harald Paganetti. Range uncertainties in proton therapy and the role of monte carlo simulations. Physics in Medicine & Biology, 57(11):R99, 2012.

List of Publications

Peer-reviewed articles

Sui, Z., Palaniappan, P., Paganelli, C., Kurz, C., Landry, G. and Riboldi, M., 2024. Imaging error reduction in radial cine-MRI with deep learning-based intra-frame motion compensation. Physics in Medicine & Biology, 69(22), p.225011.

Sui, Z., Palaniappan, P., Brenner, J., Paganelli, C., Kurz, C., Landry, G. and Riboldi, M., 2024. Intra-frame motion deterioration effects and deep-learning-based compensation in MR-guided radiotherapy. Medical Physics, 51(3), pp.1899-1917.

Lombardo, E., Velezmoro, L., Marschner, S.N., Rabe, M., Tejero, C., Papadopoulou, C.I., **Sui, Z.**, Reiner, M., Corradini, S., Belka, C. and Kurz, C., 2024. Patient-specific deep learning tracking framework for real-time 2D target localization in MRI-guided radiotherapy. International Journal of Radiation Oncology* Biology* Physics.

Conferences

Sui, Z., Palaniappan, P., Paganelli, C., Kurz, C., Landry, G. and Riboldi, M., 2024. PP01. 09 AI-BASED IMAGING ERROR REDUCTION IN RADIAL CINE-MRI: A PROOF OF CONCEPT. Physica Medica, 125, p.103580.

Sui, Z., et al. Compensation of Intra-frame Motion Deterioration Effects in MR-guided Radiotherapy. AAPM 65th Annual Meeting & Exhibition, Houston, TX, July 23-27, 2023.

Rädler, M., Meyer, S., Brenner, J., **Sui, Z.**, Landry, G., Dedes, G., Gianoli, C., Parodi, K., Riboldi, M. and Palaniappan, P., 2023, November. Investigating the benefit of scattering in 2D-3D rigid registration using a single proton radiography. In 2023 IEEE Nuclear Science Symposium, Medical Imaging Conference and International Symposium on Room-Temperature Semiconductor Detectors (NSS MIC RTSD) (pp. 1-1). IEEE.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Dr. Marco Riboldi, for his invaluable guidance and support throughout my doctoral research at LMU. Approximately four years ago, thanks to the opportunity provided by Prof. Riboldi, I left my familiar surroundings and relocated to Munich during the pandemic, embarking on a journey that has become one of the most memorable experiences of my life. I deeply admire the German term *Doktorvater*. Being far from home in a country with a completely different culture and language, I consider myself incredibly fortunate to have had such a mentor—someone who introduced me to the interdisciplinary field of Medical Physics and supported me with patience, trust, and dedication—not only academically, but also emotionally and financially. Thank you for the countless in-depth discussions and insightful advice on both technical and non-technical matters, for encouraging me to pursue my ideas, and for all the opportunities you provided, ranging from summer schools and academic conferences to side projects and many other enriching experiences. Thank you for all the wonderful activities and moments we shared. The knowledge I have gained from you, as well as the warmth and care you extended to me, have profoundly shaped both my academic and personal growth.

I would like to thank Prof. Dr. Chiara Paganelli for serving as the second referee of my thesis, and for her significant research on the 4D MRI digital phantom, which provided the fundamental materials and validation tools for my research project. I deeply appreciate her kind support and encouragement during our discussions. I would also like to express my gratitude to Prof. Bernhard Mayer, Prof. Mark Wenig and PD Dr. Torsten Enßlin for kindly accepting the invitation to serve on my defense committee.

I am sincerely thankful to Prof. Dr. Guillaume Landry and Dr. Christopher Kurz for their help, scientific guidance, and input over the years, as well as for spending time with me to conduct the MR-Linac latency measurement experiment, which became the driving force behind my research direction. I would also like to express my gratitude to Dr. Moritz Rabe for introducing me to the ViewRay MRidian MR-Linac for the first time, to Elia Lombardo for the fruitful discussions on motion prediction, and to the other colleagues in the research group of the Department of Radiation Oncology at the University Hospital, LMU Munich. I am grateful for having joined their U3 physics seminar at the beginning of my research, which was truly inspiring and provided valuable insights. Additionally, I would like to thank the Research Training Group GRK2274 for giving me the opportunity to participate in various seminars, retreats, and courses, which broadened my perspective on other areas of Medical Physics beyond my own field of study.

It has been my great pleasure to work in the Chair of Medical Physics, becoming a member of this wonderful team with a culture that emphasizes well-being and mutual respect. I would like to sincerely thank all my colleagues who have made this journey both enjoyable and incredibly meaningful. From our chair seminars and retreats to our hiking trips, Christmas events, and coffee breaks, these moments have left me with lasting and heartwarming memories. I am especially grateful to our group leader, Prof. Dr. Katia Parodi. I appreciate her dedication and contributions to both the field of Medical Physics and our team. I would also like to thank Prof. Dr. Peter Thirolf for organizing the Garchinger Maier-Leibnitz-Kolloquium and for each warm and enlightening conversation with him. I am thankful to Romy Knab, Andrea Leinthaler, Petra Glier, and Eileen Helm for their administrative support. Special thanks to Dr. Felix Rauscher for his technical assistance and for sharing his expertise in computer science. I am also very grateful to Leonard Doyle for patiently reviewing the Zusammenfassung and providing highly constructive comments. Additionally, I would like to thank my fellow Chinese colleagues in the group for their support, friendship, and the sense of belonging they brought throughout my PhD journey. In particular, I would like to express my heartfelt gratitude to Tianxue Du—our daily lunch breaks, filled with sharing, encouragement, and mutual support, made him the kind of colleague I had always dreamed of.

My sincere appreciation also goes to the members of AG Riboldi. I would like to particularly thank my office mate Dr. Prasannakumar Palaniappan for all our discussions, food and jokes. Thank you for the opportunity to let me participate in the side project of deformable 2D-3D image registration of pCT. I am deeply thankful to Jakob Brenner for the collaborations in XCAT phantom creation. Additionally, I would also like to thank Nawal Alqethami, Tobias Fisher, Dr. Martin Rädler, Francesca Vacca, Marie Gold, Jana Spiering and Yang Gao for their valuable contributions and good times.

I would like to express my sincere gratitude to Prof. Dr. Olaf Dietrich, who took my question about rephrasing time of the MR slice-selective gradient so seriously—spending a great deal of time, even running simulations and creating a dedicated webpage. I truly enjoyed our exchange and deeply appreciate your rigorous and pragmatic academic attitude, as well as the script you kindly shared. My gratitude also extends to Tengda Zhang and Dr. Yiling Wang, who generously shared their valuable and incredibly helpful job-hunting advice with me.

This work was funded by the China Scholarship Council within the joint LMU-CSC program. I am deeply grateful to my great and ever-evolving homeland, China—may it continue to prosper and thrive. I also extend my sincere thanks to my beloved LMU and the city of Munich for providing a supportive academic environment and a welcoming home throughout this journey.

On a more personal note, I would like to thank my best friends in Munich, *Chenying Liu*, *Fan Fan*, *Lianren He* and *Sining Chen*, who, in their infinite wisdom, insisted on having their names displayed here in a special font. Our wonderful group of five has created most of the joy and significant moments in my life over these years. I would also like to express my gratitude to all the teachers who have guided me on my educational journey, as well as to my family and friends who have crossed my path, even though I cannot name everyone individually here.

I would like to express my deepest gratitude to my father, Li Sui, and my mother, Hongmei Guo, for their unwavering love and everything they have done for me. Finally, thank you, Dr. Lianren He, for always being by my side and for your care, warmth, and support throughout this journey.