| | |
|---|---|
| **Titre:**<br>Title: | Image-Based Analysis and Modelling of Respiratory Motion Using Deep Learning Techniques |
| **Auteur:**<br>Author: | Liset Vazquez Romaguera |
| **Date:** | 2021 |
| **Type:** | Mémoire ou thèse / Dissertation or Thesis |
| **Référence:**<br>Citation: | Vazquez Romaguera, L. (2021). Image-Based Analysis and Modelling of Respiratory Motion Using Deep Learning Techniques [Ph.D. thesis, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/9915/ |

| | |
|---|---|
| **URL de PolyPublie:**<br>PolyPublie URL: | https://publications.polymtl.ca/9915/ |
| **Directeurs de recherche:**<br>Advisors: | Samuel Kadoury, & Jean-François Carrier |
| **Programme:**<br>Program: | Génie biomédical |

**POLYTECHNIQUE MONTRÉAL**
affiliée à l'Université de Montréal

# Image-based analysis and modelling of respiratory motion using deep learning techniques

**LISET VAZQUEZ ROMAGUERA**
Institut de génie biomédical

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*

Génie biomédical

Décembre 2021

**POLYTECHNIQUE MONTRÉAL**
affiliée à l'Université de Montréal

Cette thèse intitulée :

**Image-based analysis and modelling of respiratory motion using deep learning techniques**

présentée par **Liset VAZQUEZ ROMAGUERA**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

**Guillaume-Alexandre BILODEAU**, président
**Samuel KADOURY**, membre et directeur de recherche
**Jean-François CARRIER**, membre et codirecteur de recherche
**Benjamin DE LEENER**, membre
**Nikos PARAGIOS**, membre externe

## DEDICATION

*To my husband, my loving parents and sister, who every day gave me the fortitude to continue my doctoral journey.*

## ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my director, Samuel Kadoury, who gave me the opportunity to pursue my doctoral studies at his lab, and guided me throughout the research.

I would like to thank my codirector, Jean-François Carrier, who introduced me to the clinical knowledge I needed to start this research project and was always supportive.

I would like to convey my gratitude to all members of my Ph.D. committee for accepting to be part of the jury.

I would also like to thank NSERC, Elekta, and the TransMedTech Institute for their financial aid. Beyond the financial aspect, the TransMedTech Institute and the Advanced Development team of Elekta in Montreal also contributed to my professional formation.

I am deeply thankful to all my colleagues at the MedICAL laboratory, especially those with whom I had a close collaboration. Thank you, Rosalie and Nils, for your guidance and support while I was giving the first steps in this project. Special thanks to Tal, for his help and serenity while we sorted the research challenges.

My special gratitude goes to my family, for their love, support, and patience during my studies.

# RÉSUMÉ

En Amérique du Nord, les tumeurs primitives du foie et les métastases associées représentent la deuxième cause de mortalité liée au cancer, causant plus de 600 000 décès chaque année. Pour les patients dans les stades intermédiaires et avancés, la radiothérapie externe est souvent utilisée pour traiter et contrôler la progression de la maladie. Son objectif est de fournir suffisamment d'irradiation aux cellules cancéreuses afin d'endommager le matériel génétique de celles-ci. Par conséquent, lors de l'administration de la dose, l'objectif est d'obtenir une cible presque statique dans le champ de vision du faisceau lorsque celui-ci est actif. Cependant, le mouvement des organes respiratoires est un facteur de complication dans le traitement des tumeurs. Par conséquent, la localisation précise de la cible est l'un des principaux défis au cours de la procédure. Pour garantir une dose précise, des stratégies de gestion de la respiration sont utilisées afin d'optimiser l'administration du rayonnement au site tumoral. À cet égard, la proposition de nouvelles méthodes pour imager le mouvement de l'organe en respiration libre permettra d'évaluer quantitativement les déformations induites par la respiration.

Bien que l'imagerie 3D soit possible, les temps d'acquisition et de traitement ne sont pas compatibles avec le suivi en temps réel. Pour cette raison, les approches permettant d'obtenir de l'information volumétrique à partir de coupes 2D sont considérées comme des solutions potentielles pour améliorer l'efficacité du traitement. De plus, pour ajuster la dose conforme, il est nécessaire de prédire la trajectoire du mouvement de la cible afin de compenser les latences du système. Les développements technologiques récents ont permis la livraison simultanée de faisceaux d'irradiation et l'acquisition d'images. Les images acquises en temps réel permettent une surveillance de la cible dans le plan imagé. Elles représentent également des signaux substituts internes, ce qui les rend compatibles avec des modèles de mouvement.

L'objectif de cette thèse est de fournir un ensemble d'outils pour analyser et modéliser le mouvement respiratoire sur la base d'images généralement acquises par le flux de travail clinique. La méthodologie adoptée a conduit à répondre à trois objectifs de recherche spécifiques. Le premier objectif vise à développer une méthodologie de réordonnancement automatique de tranches pour reconstruire des volumes 4D à partir d'images IRM sans navigateur. Le second cherche à trouver de nouvelles solutions pour tirer parti de l'intelligence artificielle pour la tâche de modélisation du mouvement. Le dernier objectif vise à concevoir un prédicteur temporel basé sur l'image qui peut être intégré aux modèles de mouvement pour permettre la génération de volumes à l'avance.

La première contribution vise plus particulièrement la reconstruction automatique de volumes IRM en 4D à partir d'acquisitions cinématiques multicoupes sans navigateur. Le processus de triage des tranches pour conformer les volumes temporels est une tâche complexe. C'est encore plus difficile lorsque les signaux du navigateur ne sont pas disponibles. Nous proposons une méthodologie pour dériver des signaux pseudo-navigateurs à partir des séries dynamiques acquises à chaque position de coupe anatomique. Elle repose principalement sur le recalage d'images déformable. Par la suite, un signal unidimensionnel est dérivé d'une analyse statistique de la composante principale du mouvement. Ce signal respiratoire est ensuite traité pour identifier automatiquement un point temporel correspondant à l'état de fin d'expiration. De plus, nous avons conçu une approche basée sur des graphes pour l'empilement de tranches où les images 2D représentent les sommets. Les arêtes des graphes sont pondérées selon des mesures de similarité basées à la fois sur les informations de pixel et du mouvement. La détection automatique du temps de fin d'expiration et l'extraction automatique du pseudo-navigateur permettent à la méthode d'opérer sur des données à haute résolution spatiale et temporelle qui capturent plusieurs cycles respiratoires, permettant des études de variabilité inter-cycles. Les mesures quantitatives et qualitatives montrent une meilleure cohérence spatio-temporelle avec la méthode proposée. Comparée à des techniques similaires, qui supposent un schéma respiratoire régulier, notre méthode est capable de faire face à la respiration irrégulière et aux courtes apnées chez les sujets.

La deuxième contribution vise à proposer des solutions basées sur l'apprentissage profond pour la modélisation du mouvement respiratoire d'un point de vue à la fois déterministe et probabiliste. Les deux approches reposent sur la réduction de dimensionnalité pour associer des observations partielles à des déformations de grande dimension. Plus précisément, nous proposons l'auto-encodage convolutif comme base pour effectuer la tâche de modélisation. Le premier modèle développé associe les images substituts aux déformations de dimension réduite en minimisant la distance L2 entre les deux représentations latentes. En revanche, le second modèle utilise les images afin de conditionner une distribution probabiliste sur les déformations à chaque phase respiratoire. Les orientations sagittale et coronale ont été explorées dans nos expériences. Contrairement aux modèles de mouvement statistiques traditionnels, qui nécessitent de trouver des correspondances entre les sujets, nos méthodes exploitent la forte capacité de généralisation des réseaux profonds pour identifier ces correspondances dans un ensemble de données d'une population. Par conséquent, ces travaux présentent un nouveau paradigme pour aborder la tâche de modélisation du mouvement respiratoire. De plus, ils offrent des avantages en termes d'interprétation et de personnalisation des modèles.

Nos études révèlent que, quelle que soit la modalité d'imagerie, dans l'espace latent, les données sont regroupées en fonction de leur position dans le cycle respiratoire. D'autre part,

les modèles peuvent être facilement personnalisés à de nouveaux sujets en ajustant leurs paramètres. Fait important, étant donné que le temps d'inférence est de l'ordre de quelques millisecondes, ces modèles sont applicables en temps réel. L'analyse expérimentale sur des ensembles de données réels a montré que le modèle peut être appliqué sur des sujets exclus de l'ensemble de données d'entraînement, offrant une précision cliniquement pertinente. L'approche déterministe permet un suivi de cible en 3D à partir de tranches en 2D avec des erreurs moyennes de 2.4 mm et 5.2 mm pour des cas tests d'ensembles de données IRM et US, tandis que la variante probabiliste du modèle a obtenu une erreur moyenne de 1.67 mm et 2.17 mm dans ces mêmes ensembles de données.

Enfin, la troisième contribution propose des mécanismes prédictifs temporels pour la représentation et la génération d'images futures. Cette étape est fondamentale pour une administration et une planification précise de la dose. Cependant, elle n'est pas exempte d'obstacles tels que la prédiction à partir de dynamiques limitées ainsi que la grande dimensionnalité inhérente aux déformations complexes. Le premier modèle développé exploite les représentations de caractéristiques à plusieurs échelles et apprend à les extrapoler dans le temps à l'aide de couches récurrentes convolutives. Contrairement aux approches connexes qui tentent de régresser les valeurs dans le domaine des pixels, nous tirons parti des transformations spatiales pour relever ce défi et éviter la synthèse directe de pixel. Ce modèle est capable de prédire les positions des vaisseaux sanguins dans la prochaine image temporelle avec une précision médiane (écart interquartile) de 0.45(0.55) mm, 0.45(0.74) mm et 0.28(0.58) mm dans les ensembles de données IRM, US et CT, respectivement.

Dans cette même avenue de recherche, nous étudions également les structures d'attention de produits scalaires à têtes multiples, qui ont été initialement proposées pour le traitement du langage naturel. Ces modèles projettent linéairement l'entrée sur un ensemble de vecteurs, à savoir des requêtes, des clés et des valeurs. Contrairement à la structure originale, qui utilise le langage cible comme requêtes dans la partie décodante, nous proposons de prédire la future représentation à partir d'une séquence d'images en apprenant les requêtes. De plus, nous exploitons les images futures, disponibles lors de l'entraînement du modèle, pour calculer une distribution à priori. Cette connaissance préalable agit comme régularisateur pour l'apprentissage des requêtes. La méthode proposée est capable de prédire les déformations futures avec une erreur géométrique moyenne de $1.2 \pm 0.7$ mm dans l'ensemble de données IRM.

De plus, nous introduisons une nouvelle approche pour améliorer le suivi local. Étant donné que les méthodes de détection locales sont généralement plus précises que les prédictions de déformations denses globales, nous proposons de tirer parti des modèles de mouvement

précédemment développés pour raffiner les champs de déformation à l'intérieur d'une région d'intérêt présélectionnée autour de la cible. Cela signifie qu'au lieu de compter uniquement sur le champ de déformation global, nous l'utilisons pour améliorer le suivi de la cible locale. De plus, nous utilisons les codes latents du modèle de mouvement pour créer une carte d'attention sur les champs de déformation grossiers. Ce module de suivi est indépendant du modèle de mouvement et du prédicteur temporel. Les résultats expérimentaux révèlent qu'il peut réduire l'erreur du modèle de mouvement d'environ 63%.

Ce projet de recherche nous a permis d'étudier l'utilisation des réseaux de neurones profonds pour la modélisation des déformations de grande dimension dans un espace latent et de les relier à des observations partielles. De plus, il a introduit le premier modèle basé sur une population de sujets utilisant des réseaux génératifs profonds appliqués au suivi des mouvements respiratoires. Cette recherche a démontré que les modèles proposés peuvent également être personnalisés, les rendant plus adaptés aux caractéristiques uniques du patient. En résumé, cet ensemble de méthodes de compensation de mouvement devrait avoir un impact sur la prochaine génération d'appareils de radiothérapie guidée par l'image et devenir un élément important pour l'optimisation du traitement.

## ABSTRACT

In North America, primary liver tumor and associated metastasis represent the second most common cause of cancer-related mortality, causing more than 600,000 deaths each year. For both intermediate and late stages, external beam radiotherapy is often used to treat and control disease progression. Its goal is to deliver enough radiation to damage the genetic material of cancerous cells. Therefore, during dose delivery, the aim is to obtain a possibly static target in the beam's eye view whenever the beam is on. However, respiratory organ motion is a complicating factor in tumour treatment. Consequently, accurate target localization is one of the main challenges during the procedure. To ensure an accurate dose, respiration management strategies are required to optimize the radiation delivery to the tumor site. In this respect, proposing new methods for imaging the temporal dynamic of the organ during free-breathing will allow the quantitative assessment of respiratory-induced deformations.

Although 3D imaging is possible, the acquisition and processing times are not compatible with real-time monitoring. For this reason, approaches to obtain volumetric information from 2D slices are considered potential solutions for improving the treatment efficiency. Furthermore, to adjust the conformal dose, it is necessary to predict the target motion trajectory in advance in order to compensate for the system latencies. Recent technological developments have enabled simultaneous beam delivery and image acquisition. The real-time image acquisitions allow for in-plane target monitoring. At the same time, they act as internal surrogates, making them suitable to drive motion models.

The focus of this thesis is to provide a set of tools for analyzing and modelling the respiratory motion on the basis of images that are typically collected through the clinical workflow. The adopted methodology led to addressing three specific research objectives. The first objective is aimed at developing an automatic slice reordering methodology to construct 4D volumes from navigator-less MR images. The second one seeks to find novel solutions to leverage artificial intelligence for the motion modelling task. The last objective is aimed at designing an image-based temporal predictor that can be integrated into the motion models to enable future volume generation.

The first contribution aims more specifically at automatic 4D MR volume construction from navigator-less multi-slice cine acquisitions. The slice sorting process to build temporal volumes is a challenging task. It is even more difficult if navigator signals are not available. We propose a methodology to derive pseudo navigator signals from the dynamic series acquired at each anatomical slice position. It relies primarily on deformable image registration. Sub-

sequently, a uni-dimensional signal is derived from a statistical analysis of the main motion component. This respiratory signal is then processed to automatically identify a time point corresponding to the end-exhale state. Furthermore, we designed a graph-based approach for slice stacking where 2D images represent the vertices. The edges of the graphs are weighted according to similarity measures based on both pixel and motion information. The automatic end-exhale time detection and the automatic pseudo navigator extraction allow the method to work on high spatial and temporal resolution data that capture several respiratory cycles, enabling inter-cycle variability studies. Both quantitative and qualitative measures show improved spatiotemporal consistency with the proposed method. Compared to similar techniques, which assume a regular respiratory pattern, our method is able to cope with irregular breathing and small apneas of the volunteers.

The second contribution intends to establish deep learning-based solutions for respiratory motion modelling from both deterministic and probabilistic points of view. Both approaches rely on dimensionality reduction to associate partial observations with high-dimensional deformations. Specifically, we propose convolutional autoencoding as a backbone for the modeling task.

The first developed model associates the surrogate images to the low dimensional deformations by minimizing the L2 distance between both latent representations. In contrast, the second model uses the images to condition a probabilistic distribution over the deformations at each respiratory phase. Sagittal and coronal orientations were explored in our experiments. Unlike traditional statistical motion models, which require finding inter-subject correspondences, our methods exploit the strong generalization capability of deep networks to find patterns across a population dataset. Hence, these works present a novel paradigm to approach the respiratory motion modelling task. Additionally, they offer other advantages in terms of model interpretability and personalization.

Our studies reveal that, regardless of the imaging modality, data points in the latent space are clustered according to their position within the respiratory cycle. On the other hand, the models can be easily personalized to new subjects by fine-tuning their weights once created. Importantly, since the inference time is on the order of a few milliseconds, these models are real-time applicable. Experimental analysis on real datasets showed that the model can be applied on unseen subjects to yield a clinically relevant accuracy. The deterministic approach enables 3D target tracking from single-view slices with mean landmark errors of 2.4 mm and 5.2 mm in unseen cases of MRI and US datasets, while the probabilistic variant obtained a mean error of 1.67 mm and 2.17 mm in these datasets.

Finally, the third contribution proposes temporal predictive mechanisms for future image

representation and frame generation. This step is fundamental for accurate dose delivery and planning. However, it is not exempt from hurdles, such as the prediction from limited dynamics and the high-dimensionality inherent to complex deformations. The first developed model leverages feature representations at multiple scales and learns to extrapolate them through time using convolutional recurrent layers. In contrast to related approaches, which attempt to regress values in the pixel domain, we leverage spatial transformations to tackle this challenge and avoid direct pixel synthesis. This model is able to predict vessel positions in the next temporal image with a median accuracy (interquartile range) of 0.45(0.55) mm, 0.45(0.74) mm and 0.28(0.58) mm in MRI, US and CT datasets, respectively.

In this same research line, we also investigate multi-head dot-product attention structures, which were originally proposed for natural language processing. These models linearly project the input to a set of vectors, namely, queries, keys and values. Unlike the original structure, which uses the target language as queries in the decoding part, we propose to predict future representation from an image sequence by learning the queries. Furthermore, we leverage future frames, available during model training, to compute a prior distribution. This prior knowledge acts as a regularizer for learning the queries. The proposed method is able to predict future deformations with a mean geometrical error of $1.2 \pm 0.7$ mm in the MRI dataset.

Additionally, we introduce a novel approach to improve local tracking. Since local detection methods are generally more accurate than global dense deformation predictions, we propose to leverage the previously developed motion models to refine the deformation fields over a pre-selected region of interest around the target. This means that, instead of relying solely on the global DVF, we use it to enhance the local target tracking. Besides, we use the latent codes of the motion model to compute an attention map over the coarse deformation fields. This tracking module is agnostic to the motion model and the temporal predictor. Experimental results reveal that it can reduce the motion model error by approximately 63%.

This research project enabled us to investigate whether deep neural networks would be a feasible option to model high-dimensional deformations in a latent space and to relate them to partial observations. Moreover, it introduced the first population-based model using deep generative networks applied to respiratory motion tracking. This research demonstrated that the proposed models could also be personalized, making them better suited to the patient's characteristics. In summary, this set of motion compensation methods is expected to impact the next generation of image-guided radiotherapy and become an important component for treatment optimization.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF SYMBOLS AND ACRONYMS

| | |
|---|---|
| US | Ultrasound |
| MRI | Magnetic Resonance Imaging |
| CBCT | Cone Beam Computed Tomography |
| EBRT | External Beam Radiation Therapy |
| SI | Superior-Inferior |
| AP | Anterior-Posterior |
| LR | Left-Right |
| HCC | Hepatocellular carcinoma |
| DNA | Deoxyribonucleic acid |
| LINAC | Linear Accelerator |
| MR-LINAC | Magnetic Resonance Linear Accelerator |
| CT | Computed Tomography |
| PET | Positron Emission Tomography |
| ITV | Internal target volume |
| PTV | Planning target volume |
| CTV | Clinical target volume |
| OAR | Organ at risk |
| IGRT | Image-guided Radiation Therapy |
| PCA | Principal Component Analysis |
| CNN | Convolutional Neural Network |
| RNN | Recurrent Neural Network |
| GRU | Gated Recurrent Unit |
| LSTM | Long Short-Term Memory |
| GAN | Generative Adversarial Network |
| AE | Autoencoder |
| CAE | Convolutional Autoencoder |
| VAE | Variational Autoencoder |
| CVAE | Conditional Variational Autoencoder |
| STN | Spatial Transformer Network |

# LIST OF APPENDICES

# CHAPTER 1 INTRODUCTION

According to the World Health Organization, liver cancer is the third leading cause of death, accounting for more than half a million deaths each year across the world [15]. In North America, primary liver tumor, known as hepatocellular carcinoma (HCC), and secondary liver tumor (metastasis) represent, in aggregate, the second most common cause of cancer-related mortality, causing more than 600,000 deaths each year. Liver cancer is on the rise in Canada, with 24,400 new cases and 9,300 associated deaths in 2014. In parallel, the rate of liver metastases also increases. Liver metastases develop in 45% of patients with colorectal carcinoma (CRC) and currently represent a major health challenge [16], with 22,500 new cases in 2010 and 9,300 associated deaths. Patients with untreated liver cancer have poor prognosis, with a median survival of 4-9 months and 5-year survival rate of less than 5% [17]. Depending on tumor stage, curative treatment is favored [18]. Whenever possible for early stage cancer, surgery, percutaneous radiotherapy or ablative therapy are favored treatment options while for intermediate stage disease, trans-arterial chemoembolization is considered as a first-line palliative treatment in eligible patients. However, for both intermediate and late stages, external beam radiotherapy (EBRT) remains one of the most common forms of treatment. In fact, radiation therapy is applied in more than 50% of cancer patients to treat and control disease progression [19].

External beam radiotherapy employs an external radiation source and collimators to deliver precise doses of radiation to the cancerous tissue from different orientations around the patient's body. Its goal is to deliver enough radiation to damage the genetic material of cancerous cells, thus disabling them from dividing and growing the cancerous tumor further [20]. Nonetheless, radiation is not only harmful to cancerous cells, it can also damage healthy tissue. Therefore, the dose delivery is preceded by a rigorous treatment planning process in order to define the target location and surrounding organs at risk (OAR). Indeed, OAR such as the spinal cord or the heart need to be spared to avoid complications. Hence, accurate target localisation is one of the main challenges during the procedure. In the case of abdomino-thoracic organs, such as the liver and lungs, the motion induced by the patient's free-breathing causes a complex non-rigid tissue deformation. Consequently, it is the major cause of positional uncertainties and have shown to have a large dosimetric impact [21]. Therefore, respiration management strategies are required to optimize the radiation delivery to the tumor site. Multiple techniques can be used to limit or temporarily eliminate the amplitude range of respiratory motion. However, they present certain limitations related to their reproducibility and potential physiological constraints, particularly in cancer patients

[22].

Currently, imaging is used throughout the entire clinical workflow, representing a key component in the process. At the planning stage it is used to delineate targets and organs at risk. Additionally, in the clinical routine, a new scan is typically performed before starting the treatment to properly position the patient. Intra-procedural images are also acquired and registered to prior acquisitions to verify the tumor target location. In this context, modalities such as X-ray imaging or cone beam computed tomography (CBCT) are commonly used. Currently, computed tomography (CT) is considered the standard imaging modality for treatment planning. However, its ionizing nature is a drawback, whereas Magnetic Resonance Imaging (MRI) is non-ionizing and offers better soft tissue contrast. For the study of moving organs, a natural extension of static 3D imaging consists of acquiring a series of dynamic images (i.e. time-resolved images) to capture the organ's temporal behaviour. This is referred as 4D imaging, time being the fourth dimension. The development of strategies for dynamic volumetric MRI data collection has increased interest in the scientific community and is an area in constant development. As part of this thesis, we develop methodologies to construct 4D volumes from navigator-less cine acquisitions.

The development of fast MR-sequences have enabled the real-time acquisition of 2D slices, useful for intra-operative structure monitoring. Nowadays, in many hospitals the OAR monitoring and target tracking using real-time images is part of the clinical process. However, it is well-known that tumors undergo complex 3D displacements. Although 3D imaging is possible, the acquisition and processing times are not compatible with real-time monitoring. For this reason, approaches to obtain volumetric information from 2D slices are considered as potential solutions for improving the treatment efficiency. For instance, motion models aim at estimating the motion undergone by the entire imaged anatomy, thereby proving global 3D information. Moreover, to adjust the conformal dose according to this feedback, it is required to predict in advance the target motion trajectory to account for the processing times. Motion models are also helpful during treatment planning, to assess the feasibility of the procedure and to determine the best possible approach of the target. The basic principle behind motion models is the mapping of dense 3D motion fields to surrogate signals, which are easy to acquire and can be represented in few dimensions. A motion model can be specifically designed for a given subject or can be created with data from multiple subjects. Despite the promising results of these models on patient data, they still present significant limitations. In particular, the construction of a single patient model requires the acquisition of 4D data for each new patient, which can be impractical. On the other hand, population models, which are expected to work on unseen cases, require significant efforts to be created.

Recent advancements in deep learning have opened new opportunities to formulate the motion modelling task given sufficiently large training datasets. Some previous works in computer vision have attempted to learn a joint mapping between partial views and prior 3D shapes. The advantage of deep learning approaches over conventional motion models lies in their ability to learn the patterns with very little human intervention. Specifically, unsupervised settings do not require expert-annotated data. This makes deep learning models very flexible and powerful when there is sufficient amount of training data. Also, the excellent generalization capabilities of neural networks enable learning over a population dataset and applying the knowledge to unseen subjects, which resembles traditional inter-subject motion models but with much less effort and time.

Despite these advantages, in the field of respiratory motion modelling little progress have been done beyond the traditional statistical modelling via principal component analysis. In addition, the application of recent concepts in deep learning for motion modeling has not been sufficiently studied to introduce innovative solutions in the clinical scenario. In this thesis, we approach the motion modelling and the image-based temporal prediction from a new and promising perspective by exploiting deep learning techniques. Therefore, we advance the knowledge in the field by proposing respiratory motion models, which we postulate may become an important component toward the next generation of image-guided radiotherapy. The next sections provide the main contributions and an overview of the organization of the manuscript.

## 1.1 Contributions

This thesis, which falls within the field of biomedical imaging, is aimed at developing methods for imaging, analysis, and modelling of respiratory organ motion. As mentioned previously, in the context of image-guided radiation therapy, motion compensation strategies aid the treatment planning and improve dose delivery. The main contributions of this work can be summarized as follows :

- Proposing a fully automatic self-sorting 4D MR volume construction method that ensures the temporal coherence of the results. It includes a methodology to derive a pseudo navigator signal from dynamic slice acquisition series and a graph-based approach for slice stacking. The automatic end-exhale time detection and the automatic pseudo navigator extraction allow the method to work on high spatial and temporal resolution data that capture several respiratory cycles, enabling inter-cycle variability studies. Compared to similar techniques that assume a regular respiratory pattern, this method is able to cope with irregular breathing and small apneas of the volunteers (Chapter 5).

- Presenting deep learning-based solutions for respiratory motion modelling from both deterministic and probabilistic point-of-views. The two proposed solutions are based on dimensionality reduction to relate partial observations with high-dimensional deformations. Specifically, we propose convolutional autoencoding as a backbone for the modeling task. In contrast to traditional statistical models, which requires establishing inter-subject correspondences, our methods rely on the strong generalization capability of deep networks to find patterns across a population dataset. Hence, the burden of this complex step, which often requires manual intervention and is time-consuming, is removed. It is replaced by unsupervised feature learning across population samples, which represents a significant benefit over the state-of-the-art (Chapters 6 and 7).
- Proposing temporal predictive mechanisms for future image representation and frame generation. Contrary to similar approaches, which attempt to regress values in the pixel domain, we leverage spatial transformations to tackle this challenge and avoid direct pixel generation. Moreover, we investigate multi-head attention structures with a learnable prior to learn the spatio-temporal dynamic of the images (Chapters 8 and 9).

## 1.2 Thesis structure

This thesis is composed of eleven chapters. Following this introduction, Chapter 2 provides background information that is useful to better understand the methods developed throughout this thesis and the context in which they were developed. Chapter 3 presents a critical literature review about 4D imaging techniques, image-based temporal predictive mechanisms, and motion modeling approaches, which are the main pillars of this work. The research problem, objectives and hypothesis, as well as the general methodology, are exposed in Chapter 4.

The main findings of this thesis are presented in four articles, which are included in Chapters 5, 6, 7 and 8. These articles have been published in peer reviewed scientific journals. Chapter 5 presents the first article entitled "Automatic self-gated 4D-MRI construction from free-breathing 2D acquisitions applied on liver images", published by the International Journal of Computer Assisted Radiology and Surgery. It presents an automatic weighted graph-based method designed for volume reconstruction from navigator-less cine acquisitions. Chapter 6 presents the second article entitled "Predictive online 3D target tracking with population-based generative networks for image-guided radiotherapy", which was published by the International Journal of Computer Assisted Radiology and Surgery. This paper presents the first population-based deep motion model reported in the literature. Chapter 7 presents the third article entitled "Probabilistic 4D predictive model from in-room surrogates using conditional generative networks for image-guided radiotherapy" published by the Medical Image Analysis journal. This paper describes a probabilistic formulation of the motion modelling task and demonstrates its use both in population-based and subject-specific conditions. Chapter 8 introduces the fourth article entitled "Prediction of in-plane organ deformation during free-breathing radiotherapy via discriminative spatial transformer networks", which was published by the Medical Image Analysis journal. It describes a multi-scale framework for in-plane motion prediction and tracking. Additionally, Chapter 9 presents an attention-based model to predict future representations from an input image sequence and explains the integration of a tracker module to refine motion fields in a given region of interest. Chapter 10 discusses the benefits brought by each development to the motion modelling community and the clinical considerations. Finally, Chapter 11 summarizes the findings, limitations, and suggested avenues for future work.

# CHAPTER 2 BACKGROUND

In this chapter we provide background information related to the research topic. First, we will explain the physiology of the respiratory motion, the bio-mechanics involved in the process, and the methods for motion estimation in medical images. Then, we will introduce the process of external beam radiation therapy, which is the treatment modality that this research is focusing on. Furthermore, we will describe the classical techniques used for motion management during treatment. Finally, we will present its more recent technology, i.e., the image-guided radiation therapy and the current commercially available clinical systems.

## 2.1 Physiology of respiratory motion

Respiration is a vital process of the human body where there is an exchange of the gases oxygen and carbon dioxide. This exchange is accomplished by the quasi-periodic, bio-mechanical process of breathing. It consists of an inhalation phase, during which oxygen-rich air flows into the lungs, followed by an exhalation phase, during which carbon dioxide is expelled from the lungs to outside. The combined actions of inspiration and expiration constitute the respiratory cycle. For a healthy adult at rest, the typical respiratory rate is 12–16 breaths per minute [23]. Moreover, it is an involuntary action since a person would continue to breathe despite being unconscious. Nonetheless, individuals are capable of controlling the frequency and magnitude of their respiration, as well as breath-holds, to a limited extent. Unlike cardiac motion, the respiratory motion is not rhythmic and irregular [22].

The inhalation and expulsion of air are aided by the movement of the diaphragm, a dome-shaped muscle attached to the inferior end of the lung that separates the thorax and abdomen. The pressure and volume within the lungs can be changed by the motion of the diaphragm and the ribs. During inhalation, the diaphragm contracts and pushes the contents of the abdomen in inferior direction while the volume of the lungs expands. Simultaneously, the external intercostal muscles also participate in the inhalation process. They expand the rib cage and raise the ribs upward and outward, thereby increasing the volume of the thoracic cavity. This allows air to enter the lungs, where gas exchange takes place between blood and oxygen within the capillaries [24]. During expiration, the events are opposite of those involved in the inspiration. The diaphragm and external intercostals muscles relax and return to their original position. As a result, the thoracic volume is decreased, the intrapulmonic pressure is forced out of the lungs, and the abdominal organs move in superior direction again. Figure 2.1 depicts the respiratory motion of the thorax for inhalation and exhalation. This type of

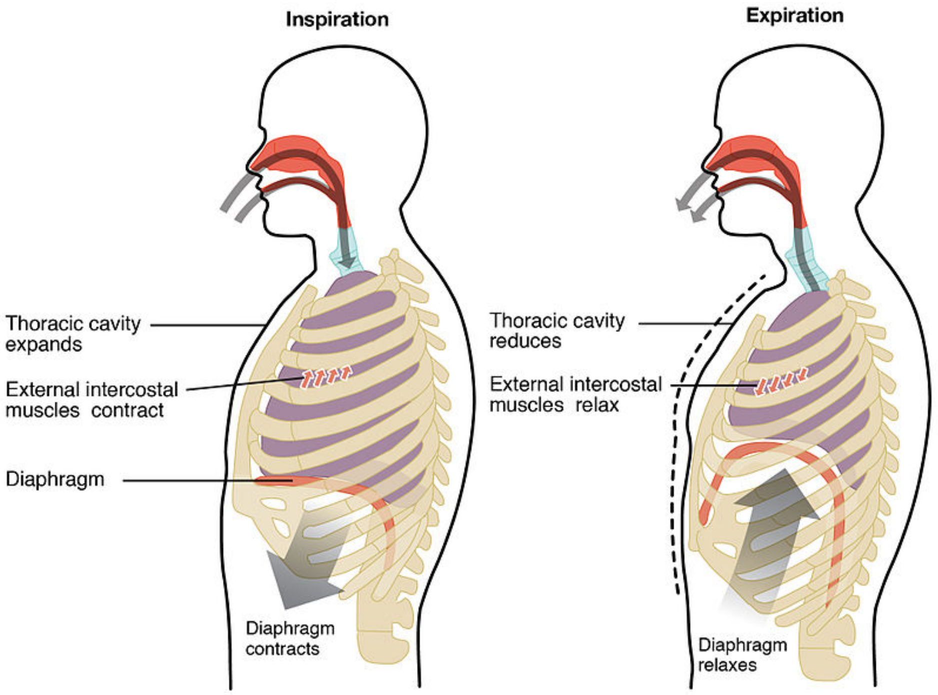


Figure 2.1 Mechanics of the respiratory motion during inhalation and exhalation within the thorax. Source [1]

breathing, which is mainly caused by the diaphragm contraction, is known as abdominal breathing. Other forms of breathing include the costal breathing, which is a rare type that occurs when negative intrapulmonary pressure is primarily achieved by contraction of the external intercostal muscle. A comprehensive overview about the physiology of respiratory motion can be found in [25].

### 2.1.1 Abdominal motion observations

The breathing process also affects other organs beyond the lungs, which are not directly involved in the respiration. For instance, the liver, esophagus, pancreas, breast, prostate, and kidneys, among other organs, are known to undergo a complex mixture of motion and deformation during free breathing. According to a report presented by the American Association of Physicists in Medicine (AAPM) [22], the average superior inferior (SI) motion for abdominal organs ranges between 10 mm and 25 mm. In the anterior posterior (AP) and left right (LR) motion planes the mean motion is less than 2 mm. These values can vary depending on the specific organ. For instance, in the liver, the SI shift ranges from 5 to 25 mm during relaxed breathing. Additionally, motion amplitudes in the AP and LR directions vary between 1 - 12 mm and 1 - 3 mm, respectively [11]. Furthermore, this motion is comprised by complex non-rigid deformations and influenced by drifts or even irregular displacements due cardiac motion, bladder filling, moving gases in the digestive tract, etc. [26].

### 2.1.2 Variability of the respiratory motion

In the literature, two types of breathing variation are distinguished: inter and intra cycle. The first describes the variation between different breathing cycles. Figures 2.2 (a) and (b) show SI motion amplitudes during several respiratory cycles. The first example corresponds to a regular breathing pattern with small variations in frequency and magnitude. In contrast, the second example, obtained from a different subject, shows irregularities in the amplitude, frequency, and shape of the breathing trajectory. It can be caused by intermittent deep and shallow breathing, coughing, emotional changes, among other factors. Therefore, depending on the subject and certain conditions, the inter-cycle variability may become non-negligible.

The other type of variation occurs within a single breathing cycle. For instance, one particular type of intra-cycle variation is the different trajectory during inhalation and exhalation. This is called hysteresis and has been described in several studies both for tumors in the liver and lungs [26]. A difference of up to 5 mm between both trajectories (inhalation and exhalation) was reported in [27]. Figure 2.2 (c) shows the typical elliptic trajectory of a liver landmark

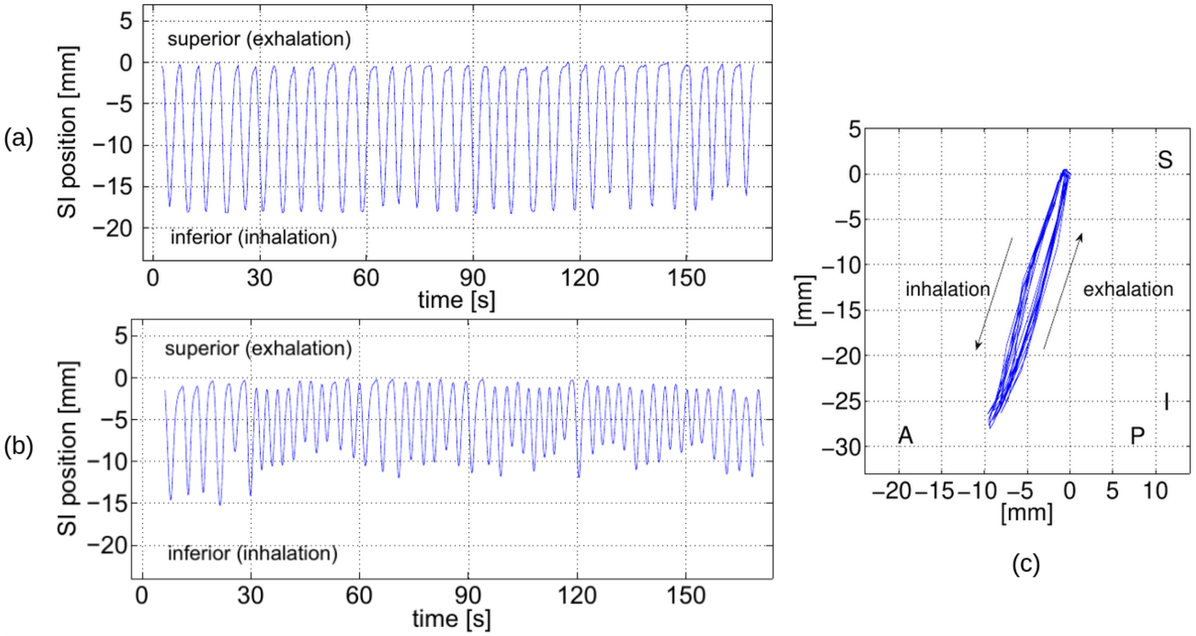


Figure 2.2 Superior-inferior and anterior-posterior motion of an exemplary point in the liver. In this sagittal view the trajectory over 20 breathing cycles exhibits hysteresis between inhalation and exhalation. Source [2]

(near the diaphragm) in the sagittal plane. In the field of radiotherapy and medical imaging, intra and inter fraction variations are the differences of motion that can be observed within one fraction and among different fractions (typically over days or weeks), respectively. These variations must be taken into account to ensure an accurate treatment delivery and/or an effective imaging.

## 2.2 Motion estimation via image registration

The goal of image registration is to align images acquired from different imaging modalities, times or subjects. Therefore, it is useful to fuse complementary information from multimodal imaging sources, to observe changes over time, or to compare the anatomy between subjects. The basic idea is to find the spatial transformation to be applied to the source image so that it is aligned with the target image. This can be found in several applications and tasks such as motion tracking, segmentation, image reconstruction, dose accumulation in EBRT, among others [28].

Generally, an algorithm for image registration involves three main components: (1) the transformation model, which defines the type of motion that is expected between the images, (2) a similarity measure, which quantifies the degree of alignment of the two images, and (3) an optimization method to find the transformation that yields the best alignment according to the similarity measure. The deformation model is determined based on the motion properties of the anatomy to register. In general, it can be classified as rigid and non-rigid. Rigid deformations are suitable for rigid objects whose movements are restricted to rotation and translation. On the other hand, non-rigid transformations constitute a large family of mappings and are suitable for describing the deformable motion of soft-tissue organs like the liver. The deformation is typically constrained by a regularization term that aims at favor specific properties in the solution. It also seeks to alleviate the difficulty associated with the ill-posedness of the problem.

Many different deformation models have been used for building motion models. Among the most commons we can cite: free-form deformations [13, 29–34], optical flow [35, 36], demons [37], biomechanical-constrained [14, 38, 39] and locally affine deformations [12, 40]. Certainly, the field of medical image registration continues to evolve rapidly. Recent efforts have been focused on using deep learning based methods, which have achieved state-of-art performances in many applications [41, 42]. Comprehensive reviews on deformable medical image registration using both traditional and deep learning techniques can be found in [43] and [28].

## 2.3 External beam radiation therapy

External beam radiation therapy (EBRT) is a common modality used to treat cancer, in which a machine directs a beam of ionising radiation through the skin to a specific part of the body where the tumor is located. The radiation beam destroys the deoxyribonucleic acid (DNA) of the malignant cells leading to cellular death using either photons (gamma-rays or x-rays), charged particles (electrons, protons or heavy ions) or uncharged particles (neutrons). These particles interact with all the tissue, including the healthy, depositing varying levels of energy as they pass through the body. Therefore, the goal is to deliver the prescription dose inside the target volume without exceeding the tolerance for the dose in normal tissues.

### 2.3.1 Respiratory motion management in radiotherapy

The motion induced by the patient's free breathing is a limiting factor during image acquisition and image-guided interventions at certain anatomical sites. During image acquisition, it can cause image artifacts thus limiting their practical utility [44]. In the context of radiation therapy, these artifacts cause distortion of the target volume and thus in the delineation of margins. Moreover, during treatment, it can produce a misalignment between the radiation beam and the moving anatomy, thereby reducing its effectiveness.

Several solutions have been proposed to reduce the impact of respiratory motion during imaging and image-guided interventions. The goal is to keep the radiation beam aligned with the target area throughout the procedure. According to [45], they can be classified into two categories: non-adaptive methods and real-time adaptive motion compensation. The former includes techniques such as using large Planning Treatment Volume (PTV) margins, abdominal compression, breath-hold, and respiratory gating. Motion encompassing methods establish large margins to cover the whole range of tumor motion. However, this increases the exposure of healthy tissues to high doses of radiation, which is undesired. Forced shallow breathing using a stereotactic body frame is an alternative method based on reducing the extent of breathing while still permitting limited normal respiration [22,46]. Intra-treatment images are essential to verify the tumor position considering the difficult reproducibility to place the compression device. Another straightforward approach is the breath-holding. In this case, the acquisition/intervention time is limited to less than 30 seconds. Some subjects are even not able to tolerate the breath-hold procedure [22]. Respiratory gating involves only acquisition/treatment during a limited portion of the respiratory cycle (e.g. end-exhalation). However, this will significantly prolong the treatment time.

Another alternative to manage respiratory motion is to move or shape the radiation beam

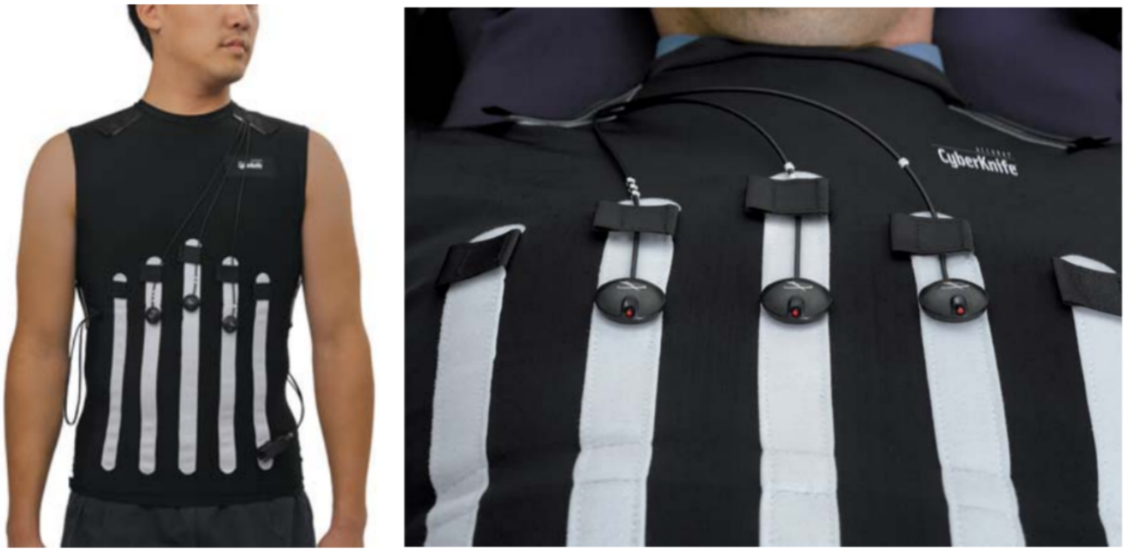


Figure 2.3 External optical device used by the Synchrony system to provide a breathing signal. Three markers, whose positions reflect the chest wall position, are attached with Velcro to a vest that the patient wears during treatment. Source [3]

dynamically as the tumor moves [3]. In this approach, known as real-time adaptive tracking, the accuracy of the dose delivery will depend on the system adapting to the moving target anatomy. Generally, the tracking methods are driven by some kind of surrogate signal to estimate the organ position. The surrogates are also referred as partial observations and can be acquired externally or internally [47]. Respiration belts and optical devices that measure the displacement of the abdominal skin are examples of external surrogate signals. For instance, in the CyberKnife Synchrony EBRT system, surrogates are acquired by measuring the displacement of the patient's abdomen or chest using optical devices (see Figure 2.3). In addition, fiducial markers are often implanted into the region of interest and tracked using an imaging device such as x-ray. This system relies in the creation of a correspondence model between the respiratory surrogate signals and the observed tumor motion.

Although the external surrogates can provide signals with high temporal resolution, in many cases there may be a low correlation with the internal organ motion either due to organ drift or varying motion patterns at different positions of the organ [24, 47]. Therefore, in clinical practice, the external surrogate is combined with low-frequency kV imaging. This allows the training and update of the correlation models, while controlling the non-therapeutic ionizing dose with respect to high-frequency fluoroscopy [48].

### 2.3.2 Respiratory motion models

Because of the limitations and drawbacks of the aforementioned techniques, there has been significant interest in the development of models for organ motion compensation during free breathing. Such models attempt to model the relationship between the motion of the organ of interest and some surrogate data. Moreover, this relationship can be used to predict future motion [49]. Generally, motion models are used when it is not possible or practical to directly measure the actual motion of interest with sufficient temporal resolution during the intended procedure.

In the literature, sometimes the terms *estimation* and *prediction* are used indistinctly. However, it is important to emphasize the distinction between the estimation of current motion and the ability to forecast future spatiotemporal displacements based on current and/or past observations. Throughout this thesis we focus rather in the later aspect, i.e. in the motion prediction, since the 4D information generated at a frequency compatible with real time applications can ultimately be integrated in a therapy planning system to improve the tumor targeting.

As stated in [50], a predictive model typically presents the following characteristics. First, its parameters should be easily adjustable to work with new subjects. Secondly, it should be robust enough for irregular motion signals, and adapt to the new breathing patterns as time evolves. Finally, it requires to quickly recover after noisy signals, such as a patient coughing. A wide variety of motion models have been proposed over the last decade mainly for lungs and liver. A comprehensive review about the current state-of-the-art is presented in section 3.3.

## 2.4 Image-guided radiotherapy

Image guidance during radiotherapy planning and treatment delivery provides essential information on target and organ locations, as well as decreases geometrical uncertainties caused by setup, breathing motion, dose-response changes, among other factors [51]. Although CBCT is the clinical standard for IGRT, the poor soft-tissue contrast and the additional ionizing dose remain as drawbacks. Ultrasound is an alternative non-ionising modality that offers high temporal resolution. However, some tumors are not visible and it completely fails for anything hidden behind absorbing structures like the ribs or inside the lung. Magnetic resonance imaging possesses superb contrast, radiation-free imaging and high temporal resolution with fast sequences [4]. Therefore, there have been significant research and commercial efforts to integrate this imaging modality into the treatment delivery devices.

Current technologies integrating MRI-guidance enable online adaptive radiotherapy delivery and imaging simultaneously. This allows a continuous visualization of target structures and surrounding organs while the treatment is being delivered. In consequence, it provides higher treatment accuracy and improved clinical outcomes. Moreover, it reduces toxicities by sparing healthy tissue. Finally, it enables efficient workflows by providing the total dose in fewer treatment sessions, which is known as hypofractionation. These features are expected to deliver real health benefits for patients, including better disease control with fewer side effects [52].

### 2.4.1 Clinical systems

Viewray MRIdian and the Elekta-Unity are two in-room MRI-guidance systems developed by commercial entities that are approved and used for treating patients (see Figure 2.4(A)). The ViewRay was the first commercial system enabling simultaneous MR imaging and a range of external-beam radiation therapy options at the same isocenter [53]. The first treatments with this unit were performed in 2014 [54] whereas the first treatment using the Elekta-Unity was reported in 2017 [55]. As illustrated in Figure 2.4 (B), both systems may use perpendicular or in-line configurations. In the latter case, the treatment beam is oriented perpendicular to the magnetic field. Thus, the SI axis of the patient is aligned with the magnetic field, and the linear accelerator can rotate independently of the magnet and patient. A drawback of this approach is that magnetic fields applied perpendicularly to the radiation beams can affect dose deposition compared to the zero field situation, particularly for higher fields. Alternatively, the magnetic field can be parallel to the radiation beams. This could minimize the effect of the magnetic field on the dose distribution. Nonetheless, it can also cause problems in certain cases. For instance, it may increase the skin dose up to 1400% [56]. This problem can be mitigated either through optimisation of the magnetic fringe field or electron purging devices [4, 56].

Overall, Viewray MRIdian consists of 3 main components. First, a a vertically gapped horizontal solenoidal superconducting 0.35 Tesla whole-body MRI. Second, the radiation delivery system, which is a robotic 3-headed $^{60}Co$ system yielding a dose rate of 550 cGy/min. Treatment monitoring is carried-out by tracking structures which are observed in fast planar images. Specifically, with a sagittal plane at 4 frames-per-second (FPS) or with 3 parallel sagittal planes at 2 FPS using real-time non-rigid registration–based beam control. Therefore, radiation beams are only enabled if the tracked region is located within the prescribed boundary with approximately 300 ms latency. Third, the adaptive RT treatment-planning system. It is an integrated software dedicated for autocontouring, Monte Carlo dose compu-

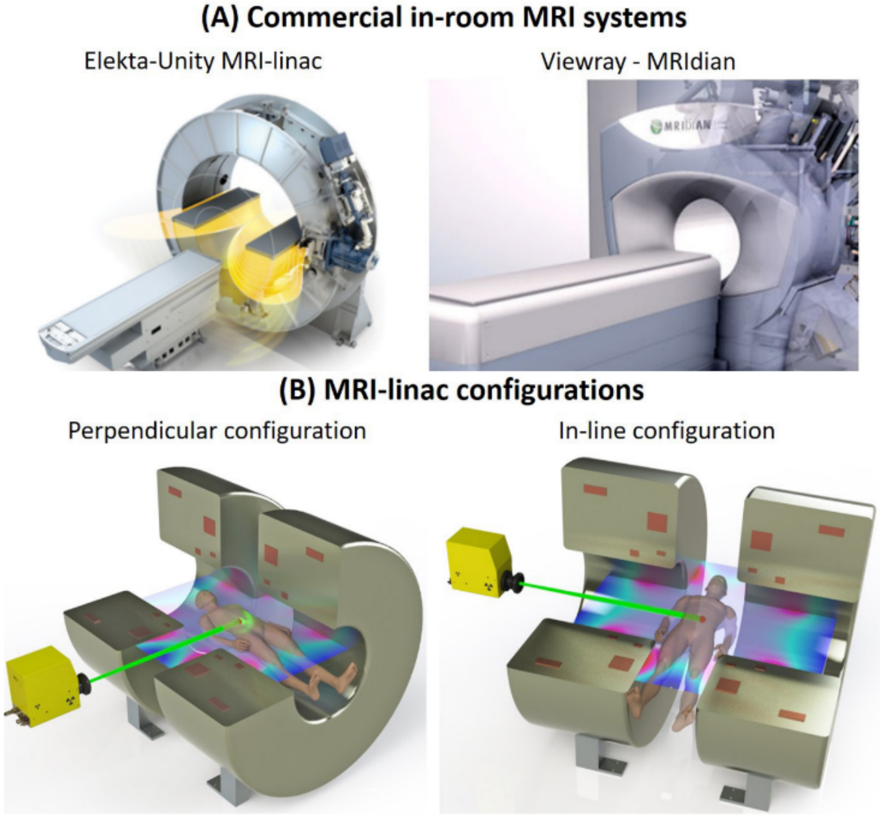

(A) Commercial in-room MRI systems
Elekta-Unity MRI-linac
Viewray - MRIdian
MRIDIAN
(B) MRI-linac configurations
Perpendicular configuration
In-line configuration

Figure 2.4 (A) Clinical systems for IGRT: Elekta-Unity MRI-linac and ViewRay MRIdian. (B) Typical configurations used for MRI-linac construction. Source [4]

tation, and conformal radiation planning. All these processes can be done within 30 seconds based on the volumetric image of the day [53]. Similarly, the Unity MR-Linac integrates a 1.5 T Philips Achieva MRI scanner with diagnostic imaging quality and a 6 MV linear accelerator. Its construction was a joint effort between the University Medical Center Utrecht, Elekta and Philips.

# CHAPTER 3 LITERATURE REVIEW

## 3.1 4D imaging of respiratory motion

Dynamic 3D imaging, also termed as 4D imaging, is crucial to quantify organ displacements and assess their mechanical functions. This has found applications in the study of diseases, treatment planning and radiation therapy [57]. The literature distinguishes two types of 4D datasets. The term respiratory-correlated 4D MRI, is used to indicate the three spatial dimensions and the respiratory phase. In contrast to time-resolved 4D images, the fourth dimension refers to time. Respiration-correlated four-dimensional X-ray computed tomography (4D-CT) was first described in 2003 [58]. Currently, it is the clinical standard for the radiation therapy workflow in the presence of respiratory motion. It allows physicians to determine the internal target volume (ITV), which is derived from the union of target volumes through the entire respiratory cycle. This information is the basis to assess 3D tumor motion and subsequently estimate the dose [59].

In 4D-CT, images are acquired and averaged over several breathing cycles. Since it is required to know the point in the respiratory cycle to which each projection corresponds, the respiratory motion is recorded using a monitoring device (e.g., a belt or an infrared (IR) marker placed on the thorax). Respiratory sorting is then performed retrospectively by correlating each projection with a point in the respiratory cycle. A prospective approach is another alternative that consists of acquiring projections at a defined point in the respiratory cycle [60–62].

The retrospective approach is much simpler since it does not require a real-time detection of the breathing state or extensive changes in the scanner software. It is based on gathering as many dynamic images as possible to cover all the possible states. Subsequently, the images are sorted based on their respiratory state. For instance, in respiratory correlated 4D CT datasets, the respiratory cycle is divided into bins (typically 10 bins) and the images are sorted according to them. Each bin is then reconstructed into a 3D dataset [58]. The bins can be determined according to the phase or the amplitude of the signal acquired by the sensors. In phase-based binning, the bins are determined by their temporal relationship to the cycle, whereas in amplitude-based binning they are determined by the amplitude of the corresponding breathing signal, which correlates with the amplitude of the diaphragm motion. Although both variants present advantages, the amplitude-based binning is more accurate than its counterpart [63]. On the other hand, there are some artifacts that commonly affect the 4D-CT datasets such as blurring, duplicate, overlapping and incomplete structures [44].

Overall, the main drawback of 4D-CT is the use of ionizing radiation and the low contrast. In certain cases, the contrast may be insufficient for tumor delineation in the abdomen.

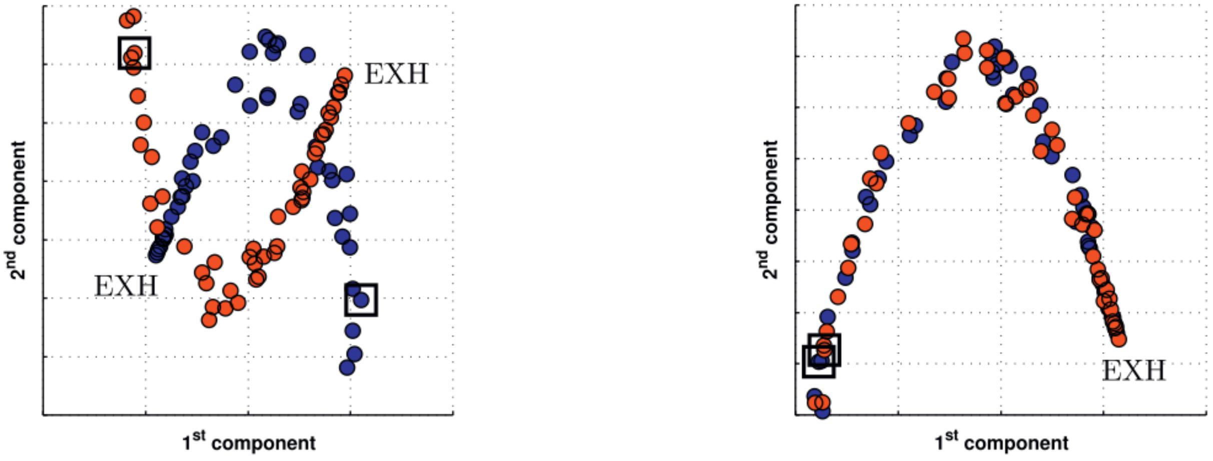
As an alternative modality, Magnetic Resonance Imaging is able to capture both anatomical and functional information. Furthermore, its high flexibility in image acquisition strategy allows excellent soft-tissue contrast with a clear distinction between tumorous tissue and organs at risk (OARs). Besides, it involves no radiation. For these reasons, its use in radiation therapy has increased considerably over the past decade. MRI aids in the delineation process, informs about functional parameters, and can be used to assess treatment response. Functional imaging, such as diffusion weighted imaging and dynamic contrast enhanced imaging, has been used successfully to discriminate between healthy and tumorous tissue. Unfortunately, currently there is no 4D technology implemented in the MRI scanners. To fill this gap, several approaches have been proposed to generate 4D-MRI datasets, which can be divided in two categories: multi-slice 2D acquisitions and 3D acquisitions [64].

### 3.1.1 Multi-slice 2D acquisitions

Multi-slice 2D acquisitions acquire the data on a slice-by-slice basis throughout the anatomy of interest. Moreover, the slices are acquired over several respiratory cycles with sufficient temporal resolution. Subsequently, the temporal images are retrospectively sorted according to the respiratory phase which is informed by navigator signals, also termed as surrogates. The surrogate signals are crucial for the subsequent slice reordering since they allow to construct dynamic volumes with a temporal consistency. They can be acquired either externally or internally.

Respiratory belts and optical tracking devices are some examples of external surrogates [65–67]. Remmert et al. [65] used an optical device as part of their 4D imaging strategy to measure the displacement of a dynamic porcine lung phantom. On the other hand, Tryggestad et al. [66] employed a belt placed around the subjects' upper abdomen to digitally encode the respiratory trace at a 50 Hz sampling rate. Liu et al. [67] used a similar external breathing monitoring device to guide the retrospective sorting. Nevertheless, these external surrogates are known to have low correlation with the internal organ motion [47], particularly in cases with irregular breathing, which causes artifacts in the sorted 4D images. In contrast, internal surrogates are often more reliable. The MR navigator echo is one of the most commonly used surrogate data. It is based on exciting a small column of magnetisation to measure the position of a region of tissue over time [47]. This approach was followed in [60] and [34] to monitor the respiratory-induced shifting of the liver and diaphragm.

Many sorting methods, whether based on external or internal surrogates, make strong as-



(a) Unaligned embeddings

(b) Aligned embeddings

Figure 3.1 Unaligned and aligned embeddings for two slice positions of one volunteer. (a) Embedding as obtained directly from Locally Linear Embedding without alignment. (b) Embedding after the alignment where nearby points correspond to similar respiratory positions. Source [5]

sumptions about the regularity of the respiratory motion and represent it with 1D signals. Von Siebenthal et al. [2] argued that parameterizing respiratory motion with one parameter neglects all residual variability and may be a coarse approximation in some cases. Hence, they proposed an interleaved scheme to acquire 2D navigator frames at a fixed anatomical position and data frames covering the imaged volume. Slice reordering was then based on assessing the similarity between 2D navigators since the respiratory state at each image slice was defined by its adjacent navigator slices. The main drawback of internal navigators is that they tend to decrease the temporal resolution of the actual acquisitions. Moreover, they may also increase the overall acquisition time [68].

Another alternative is to derive the respiratory trace using features contained in the captured images, which is known as self-sorting, self-navigation or self-guidance. Slice reordering techniques that do not rely on external or internal surrogate signals can be grouped into two main categories: machine learning [5, 69–74] and slice feature extraction-based methods [6, 57, 68, 75–80]. Manifold learning (ML) is a machine learning based technique that has shown to be useful in the analysis of motion in medical datasets [71]. In the context of slice reordering, this powerful tool has been employed to map dynamic slices from different anatomical positions into a low-dimensional space according to their respiratory phases. Some methods used to create such representation include Isomap, Locally Linear Embedding (LLE) and Laplacian Eigenmaps (LE). Before the actual slice stacking to construct the temporal volumes, the manifolds yielded at different anatomical positions are all combined within one single globally consistent embedding using Manifold Alignment (MA) techniques. Figures 3.1 (a) and (b) show the low-dimensional space before and after the alignment, respectively.

Baumgartner et al. [71] addressed the alignment of multiple manifolds obtained using LLE by overlapping groups of two. They also proposed a sparsification technique for the Gaussian inter-dataset similarity kernel calculation. Later, the authors extended their work [5] by adding a registration-based inter-dataset kernel, which incorporated knowledge of the approximate relations between adjacent slice positions. Moreover, in [72] the mathematical formulation of LLE was extended to simultaneously embed more than two datasets. The authors tackled the similarity kernel choice problem by introducing a random walk-based graph matching technique, which was used to determine such kernel. The advantage of that proposal was the global alignment of the data without prior correspondences nor comparisons between the high-dimensional data. Clough et al. [73] achieved state-of-the-art performance over the former method by introducing a novel graph-based descriptor. The main limitation of MA-based techniques is that they commonly makes assumptions about the regularity of the respiratory motion. However, there is a non-negligible residual variability that makes data discrimination difficult.

In feature extraction based methods, the derivation of a reliable respiratory signal from the acquired images is used to optimize the reordering process. Some prior works has proposed to monitor the body area to represent the breathing signal, as it typically correlates with the breathing motion [6, 76, 77] (see Figure 3.2). For instance, in [6] and [76], 2D dynamic axial images were employed for the reordering process. Nevertheless, axial planes are rarely used since abdominal motion is better appreciated in the sagittal and coronal planes. Liu et al. [77] demonstrated that sagittal slices yield more accurate 3D volume reconstructions. However, changes of body area are prone to be affected by space-dependent phase shifts.

On the other hand, some techniques compute the image similarity between contiguous slices relying on different metrics. For instance, the approaches presented in [75] and [78] are based on calculating the mutual information between slices. Unfortunately, slice reordering methods based exclusively on imaging data may not guarantee adequate temporal behavior. Alternatively, Uh et al. [79] proposed to obtain internal surrogates by applying dimensionality reduction on the dynamic slices. Specifically, they used principal component analysis to do the mapping to the low-dimensional space. However, the low-dimensional representation of the images does not always proportionally change with respiratory motion. This could be a limitation in this approach.

Another group of methods introduce several intermediate steps such as segmentation and

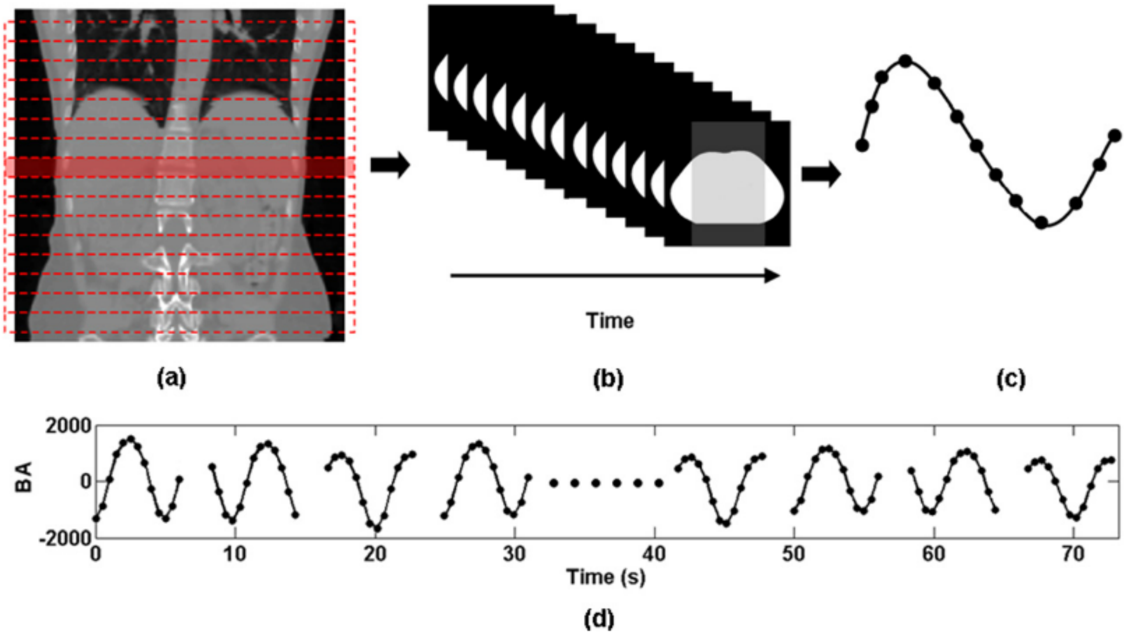


Figure 3.2 Workflow of extracting breathing signals from presorted 4D-CT images using the body area (BA) method. (a) Axial images were acquired continuously throughout the breathing cycle. (b) Body area (white area) was determined for each image. In practice, only the middle section of the image (grey area) was used for body area calculation. (c) For each image slice, an individual breathing curve was generated by plotting the body area as a function of image acquisition time. (d) The complete breathing signal is generated by plotting all individual breathing curves as a function of acquisition time. Gaps between individual breathing curves are due to couch movements during the 4D-CT scans [6].

registration to derive the respiratory trace [57,68,80,81]. The methodology proposed by [68] first computes a median intensity image from all the coronal dynamics at certain anatomical position. Then, the center of the liver dome is detected relying on a previous lung segmentation. Subsequently, the dynamic slices are rigidly registered to the median intensity image, which results in $N$ shifts per slice position ($N$ is the number of temporal slices). The set of shifts are filtered and normalized before obtaining the final self-sorting signal, which was divided into 10 bins for the sorting purpose. Tong et al. [57] proposed to construct a weighted graph and to reconstruct volumes following the shortest path. The first step is the identification of a reference image at each anatomical position. This process was done manually, which is impractical. Similarly to [68], this method relies on a lung segmentation. Thus it is limited to sequences where lungs are entirely visible. Recently, Hao et al. [82] proposed a similar technique where the breathing signal is determined based on the optical flow computed in the image time series.

### 3.1.2 3D acquisitions

In recent years, the number of 4D imaging approaches based on 3D acquisitions has increased. The 3D readout trajectories include non-Cartesian [59, 83–85], Cartesian [86, 87], and hybrid readouts [88–92]. A standard 3D readout takes multiple respiratory cycles to collect. Therefore, the acquired data is often sorted in the k-space before image reconstruction [64]. Cartesian readouts include the rotating Cartesian k-space (ROCK) trajectory introduced by Han et al. [86] and the compressed sensing partial subsampling (ESPReSSO) scheme by Küstner et al. [87].

The 3D non-Cartesian readout consists of a large number of radial projections with different polar and azimuthal angles. Sampling the k-space with the so called "golden-angle" fills the 2D k-space with radial spokes that have a relatively uniform angular distribution for any time interval. Chan et al. [93] extended this concept to the 3D space by introducing the concept of multidimensional golden means. They obtain a uniform distribution of sampled lines throughout the acquisition. This strategy achieves very high undersampling factors with benign (incoherent) image artifacts making it suitable for compressed sensing image acceleration. It is also very robust since the center of k-space is sampled by each line [64]. Deng et al. [59] proposed a radial readout of the k-space using the 2D golden means ordering and self-gating motion surrogate. This scheme yielded respiratory-resolved 3D volumes with isotropic high spatial resolution and an arbitrary number of temporal phases. Generally, the reconstruction of non-Cartesian readouts involves a 3D gridding step, which is computationally expensive. Also, oversampling the center of k-space requires time, which reduces the efficiency of the readout [64].

Hybrid readouts (radial-Cartesian) attempt to mitigate the aforementioned shortcomings. Buerger et al. [88] proposed a 3D acquisition with golden angle ordering in the phase-partition (ky–kz) plane followed by k-space sorting. The golden-radial phase encoding technique achieves high isotropic spatial resolution. Furthermore, it is robust to irregular breathing since it retrospective sorts the data in the k-space. Alternatively, some authors uses the radial stack-of-stars (SoS) technique [89–92, 94–96], where the k-space is sampled along a radial pattern in two dimensions while the third is sampled on a Cartesian grid. Stemkens et al. [92] combined a SoS acquisition with a compress sensing based reconstruction, called XD-GRASP [97, 98], to reconstruct both a DCE time series and a respiratory-correlated 4D-MRI. The combination of both techniques showed to minimize motion artifacts and to generate a respiratory-correlated 4D-MRI within a few minutes. Since a reliable respiratory signal can be obtained directly from the raw k-space data, the self-navigation property is one of its advantages.

Recent approaches have attempted to directly generate real-time volumetric images [99] or motion fields [100]. Feng et al. [99] proposed a real-time 3D MRI technique called MR SIGnature MAtching (MRSIGMA). It consists of two steps. First, an offline (non-real-time) step seeks to learn the possible 3D motion states and their association with motion signatures. A 3D readout based on golden-angle stack-of-stars was employed. The second step, which is performed online, matches real-time acquired motion signatures with the prelearned motion states. The main limitation is the adaptation to organ drifts and patient movement. On the other hand, Huttinga et al. [100] described a preliminary study aimed at recovering 3D motion fields directly from k-space data. Experimental results showcased plausible deformations with a predictive horizon of 170 ms on 5 subjects. Although this technique showed promising results, it requires further validation.

## 3.2 Image-based temporal prediction

The ability to forecast, anticipate and reason about future outcomes is a key component of computer-aided decision-making systems. With the great success of deep learning in computer vision, deep-learning-based video prediction have turned out a wide research area [8, 101–104]. Furthermore, future prediction have found applications in multiples tasks such as anticipating events [105, 106], long-term planning [107], predicting instance/semantic segmentation maps [108, 109], autonomous driving [110], anomaly detection [111] and weather forecasting [112]. Also, this area converges with other related tasks such as missing frame interpolation [101,113], action recognition [9], future trajectory prediction [114], amongst others. In the context of learning paradigms, video prediction can be defined as a self-supervised learning task as target frames are already available in the video sequence during training. Therefore, no extra labels or human supervision is needed [104].

Generally, deep architectures for video prediction are composed by an encoder, which extracts representations of prior frames, and a decoder, which generates future frames based on the extracted representations. This is quite common to all the models. On the other hand, there are distinctive elements in terms of the strategies for information processing, stochasticity, etc. In sections 3.2.2, 3.2.3, 3.2.4 and 3.2.5 we follow a similar taxonomy as in [104]. Nonetheless, it should be noted that these categories are not mutually excluding since a model could contain a combination of approaches. Besides, the terms video prediction, future frame prediction, future frame forecasting, and future frame generation are used interchangeably throughout the section.

### 3.2.1 Backbone deep learning architectures

Several deep networks have been used as core components for video prediction models, namely, recurrent neural networks, convolutional neural networks, and generative models. Recently, attention-based mechanisms, such as the so-called *Transformer* [7], emerged as a promising solution for computer vision tasks.

#### Recurrent neural networks

Recurrent neural networks (RNNs) are a type of neural network where connections between nodes form a directed graph along a temporal sequence allows information to persist. They can use their internal state (memory) to process variable sequence lengths from the inputs, which make them applicable for temporal data. Therefore, they have prevailed in many works for video prediction due to their flexibility for temporal information modeling. For instance, Ranzato et al. [115] proposed a model based on RNNs, making short-term predictions at the patch level. They divided video frames in patch clusters using k-means. A downside of this method is that results suffer of a tilling effect.

Vanilla RNNs present some limitations when dealing with long-term sequences due to the vanishing gradient issue. These shortcomings were mitigated with the introduction of more sophisticated models, such as Long Short-Term Memory (LSTM) [116] and Gated Recurrent Unit (GRU) [117]. The LSTM introduced the cell state, where information is added or removed according to gates. The GRU is a variant of the LSTM using less gates. Shi et al. [112] extended the point-wise computation performed within the LSTM to the two-dimensional space by adding the convolution operation. They applied their Convolutional LSTM (ConvLSTM) to precipitation nowcasting from radar images.

Wei et al. [118] proposed a predictive model that exploits spatial-temporal appearance information of previous frames and the inter-frame optical flow information to predict the next frame. Their model receives RGB frames observed at several time steps and corresponding optical flow maps into a two separate input streams composed by convolutional layers. The feature representation extracted by those layers were concatenated and fed to a stack of ConvLSTM to produce the RGB image for the next time step. Romaguera et al. [119] proposed a multi-scale feature extraction approach before the ConvLSTM units to predict future in-plane organ deformations in medical image sequences.

### Convolutional neural networks

Although RNNs such as LSTM and Gated Recurrent Units (GRU) have been the core component of many of these models, other works relying exclusively on feed-forward convolutional networks have also been proposed. Walker et al. [120] were presented a convolutional neural network to predict dense optical flow given a static image. Yumer et al. [121] introduced an end-to-end solution using a volumetric convolutional neural network that learned three-dimensional deformation flows. The authors proposed an architecture which took the voxelized representation of a given shape and a semantic deformation intention as input to generate a deformation flow at the output.

Watters et al. [122] introduced a convolutional neural network for learning the dynamics of a physical system from raw visual observations. Specifically, their model was composed of a visual encoder, a dynamics predictor and a state decoder. The visual encoder took a triplet of frames as input and yielded a state code which is a list of vectors, one for each object in the scene. Each of these vectors contained a distributed representation of the position and velocity of its corresponding object. Then, the dynamics predictor processed the sequence of state codes to predict a candidate state code for the next frame. Finally, the state decoder converted a state code to a physical state. In this was, they were able to infer the physical states of multiple objects and make accurate predictions about their future trajectories.

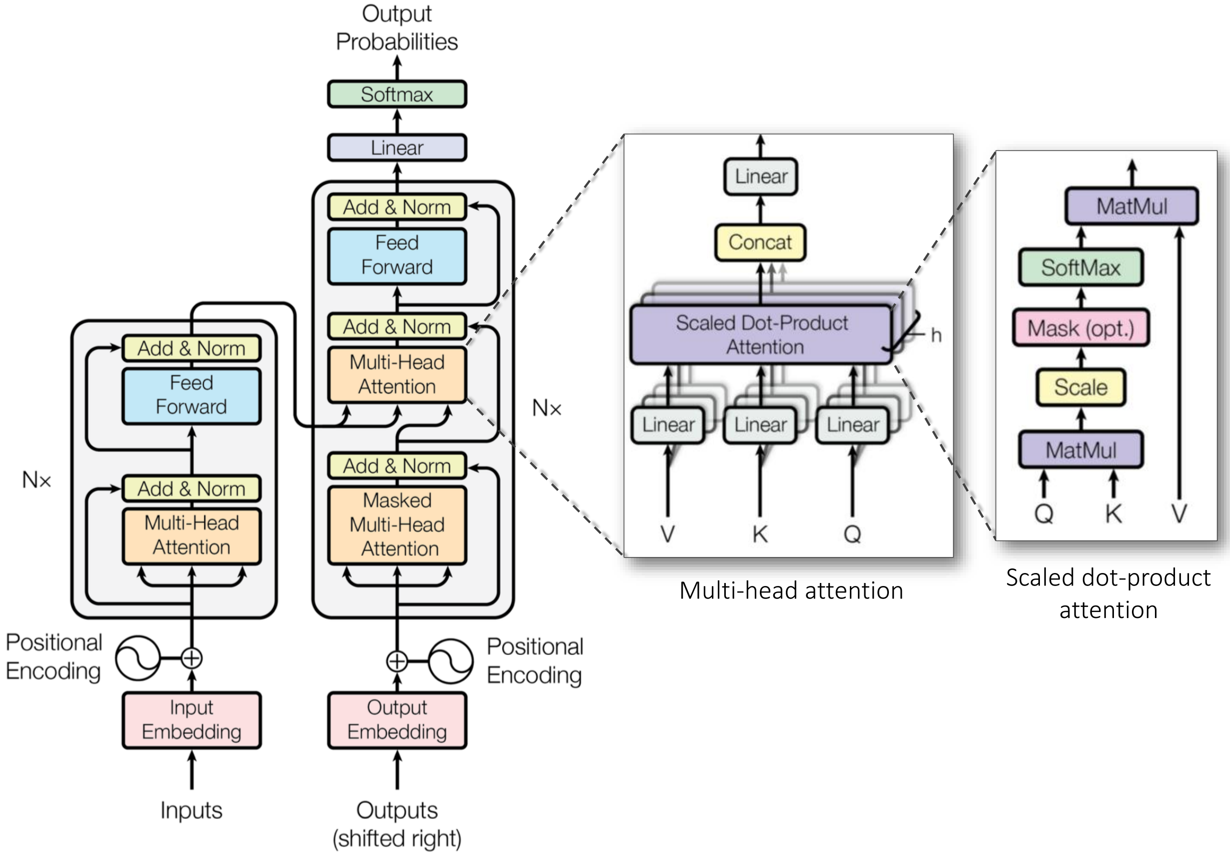
### Generative models

Variational autoencoders (VAE) and generative adversarial networks (GANs) are two popular generative models that have been used as a backbone for future frame prediction. Generative models learn the underlying distribution of each class, i.e. they capture the joint probability $p(x, y)$ or $p(x)$ in the absence of labels $y$. Moreover, given some training data, they generate new samples from the same distribution. Given input data $\sim p_{data}(x)$ and generated samples $\sim p_{model}(x)$ where, $p_{data}$ and $pmodel$ are the underlying input data and model's probability distribution respectively. The training process consists in learning a $p_{model}(x)$ similar to $p_{data}(x)$ [104]. For VAE, this is done explicitly while in GANs, it is performed implicitly by estimating a density function from the input data. The probabilistic nature of these models have been leveraged to cope with future uncertainty by generating a set of feasible predictions rather than a single outcome [104]. Since generative models have a stochastic component, the state-of-the-art on this category will be approached in Section 3.2.5.

### Dot-product multi-head attention (Transformer model)

Until a few years ago, recurrent networks used to be the prevalent solution for natural language processing (NLP). In 2017, Vaswani et al. [7] proposed a predictive model relying exclusively on dot-production attention. This means no convolutions nor recurrence was involved in the computations. This attention mechanism have shown to outperform the RNNs while introducing several advantages, namely, it helps to solve the vanishing gradient and bottleneck problems. Moreover, it increases the model interpretability. In contrast to RNNs, which process sequences recursively, Transformer can attend to complete sequences thereby learning long-range relationships and can be easily parallelized [123].

The Transformer model is composed by a stack of $N$ encoders and $N$ decoders (see Figure 3.3). Since the encoder receives the entire input sequence at once, there is no information about the ordering of each token. Therefore, positional encodings are added to the input embeddings to determine the position of each word. Each encoder block is comprised of a multi-head attention layer and a feed-forward (FF) layer. Also, residual connections and



Multi-head attention

Scaled dot-product attention

Figure 3.3 Transformer model. Source [7]

normalization are applied around these two sub-layers.

The key concept behind Transformers is the scaled dot-product attention mechanism, where the input is linearly projected to a set of queries $Q \in \mathcal{R}^{n_q \times d_q}$, keys $K \in \mathcal{R}^{n_k \times d_k}$, and values $V \in \mathcal{R}^{n_v \times d_v}$. The vector dimensionality $d_q$ equals to $d_k$, the number of keys $n_k$ equals to the number of values $n_v$. As illustrated in the rightmost part of Figure 3.3, the output of the attention layer is given by computing the weighted sum of the values, where attention scores $S \in \mathcal{R}^{n_q \times n_k}$ are calculated from the queries and key as follows:

$$A(Q, K, V) = Softmax\,(S)\,V = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.1}$$

where the $Softmax$ function is used to normalize the scaled dot-product attention scores.

The attention score determines how much focus to place on other parts of the input sentence when encoding a word at a certain position. The decoder also contains a multi-head attention layer to compute the self-attention between the decoder's inputs and a FF layer, but between them there is an additional attention layer. This layer computes a cross-attention between the decoder's inputs and encoder's outputs. Specifically, the $K - V$ pairs come from encoder, while $Q$ is derived by multi-head self-attention from the target language sentence (decoder's inputs). Thus, it helps the decoder to focus on relevant parts of the input sentence to generate the target sentence.

The success shown by Transformers in NLP tasks [124, 125] has sparked great interest in the computer vision field [126–128]. A comprehensive survey about Transformers for computer vision can be found in [123]. Girdhar et al. [129] introduced the Action Transformer model for localizing and recognizing human actions in video clips. They use a stack of convolutional layers as backbone to extract spatiotemporal features. Then the bounding boxes from a region proposal network, which localize people performing actions, are mapped into queries. Also, the clip around the person being analyzed is projected into key and values. Hence, the action recognition task was adapted to the key concepts of the Transformer architecture. The idea of using a convolutional backbone before the actual Transformer has been applied in related works to obtain a compact feature representation [127, 128]. Carion et al. [127] developed an object detection model based on Transformers, which simplified the traditional detection pipeline and achieved on par performance compared with SOTA detectors. A similar model, inspired on this work, was proposed for instance segmentation [128].

### 3.2.2 Disentangling vs. non-disentangling approaches

Depending on the strategy used to encode visual representations of the observed images, future frame prediction can be classified into non-disentangling and disentangling approaches. The former case refers whenever the encoder generates visual representations from the input frames thereby capturing spatiotemporal information using RNNs or other predictor. In this case, it is assumed that these representations contain the necessary information to make the predictions [130–134].

Alternatively, other approaches learn a disentangled representation from image sequences, i.e., they decompose the video into different components [8, 135–139]. Prior works in this direction have focused on predicting high-level semantics in a video such as action [140], events [141] and motion [142]. Villegas et al. [135] factorized the frames into a stationary part and a temporally varying component representing content and motion, respectively. This was performed using both a motion encoder and a content encoder (see Figure 3.4). The former assumes that dynamic features are captured by the difference images of two adjacent past frames. Thus, it aggregates the dynamic features from different time steps using a LSTM to form the final dynamic representation of all past frames. On the other hand, the content encoder takes the most recently observed frame as input, assuming that the appearance of two adjacent frames are close to each other. Then the features extracted in both streams are fused and passed through a decoder to predict the next frame. The multi-scale skip connections are used to boost the visual quality. Guen et al. [139] proposed a more complex kinematic motion model based on partial differential equations instead of adjacent frames difference as in [135].

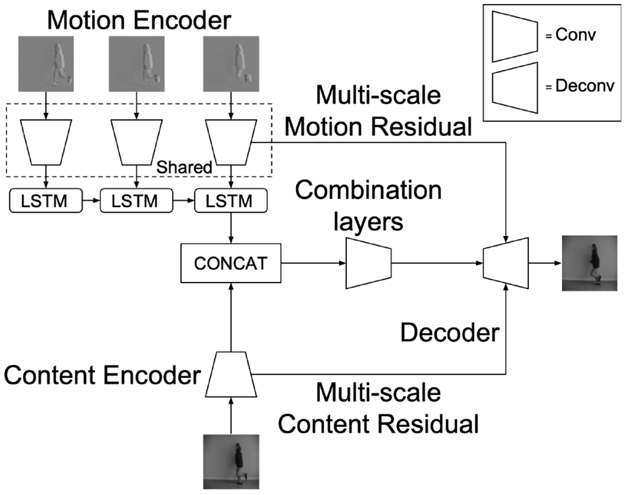


Figure 3.4 Model decomposing content and motion information. Source [8]

Similarly, Denton et al. [8] described a model where representations are disentangled into content and pose. Moreover, the poses are penalized for encoding semantic information by using a discrimination loss. Likewise, Decompositional Disentangled Predictive Auto-Encoder (DDPAE), a model proposed by Hsieh et al. [138], extracts representations of each input frame. Then, each representation is further disentangled into appearance and pose. DDPAE used an RNN with 2-dimensional recurrence. One recurrence is for the temporal modeling and the other is used to capture the dependencies between components. This model, based on variational autoencoders, enforces each component to have a low-dimensional temporal dynamic behavior.

Often the results presented in the aforementioned works are based on validations performed in object-centric datasets, with very basic or low complexity movements, for example Moving MNIST. This synthetic dataset consists of two digits moving independently in a 64 × 64 frame. However, other type of videos might be harder to directly disentangle. For instance, the disentangling concept may be more challenging in crowded scenes or in medical sequences presenting complex deformations.

### 3.2.3 Sequential vs. parallel decoding

When the predictive horizon is greater than one, the outcomes can be obtained sequentially [8, 10, 130, 131, 143, 144] or in parallel [9, 145, 146]. In the former case, the future frames are generated one by one, and each new prediction is fed back to yield the next prediction, i.e. recursively. Generally, these sequential models achieve the temporal coherence of future frames implicitly since the outputs are conditioned on previous predictions. Bengio et al. [147] proposed a a curriculum learning strategy to bridge the gap between training and inference

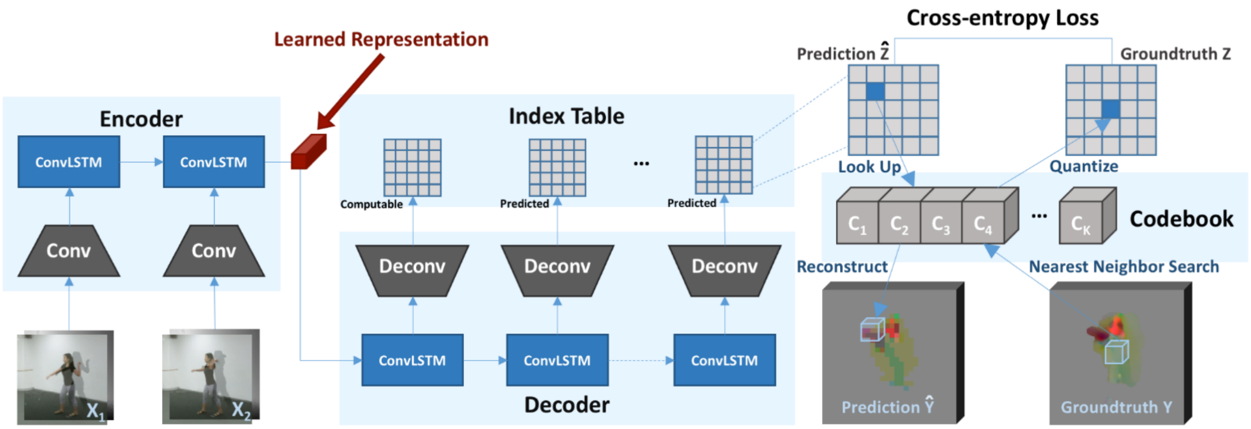


Figure 3.5 LSTM encoder-decoder for motion prediction. Source [9]

for sequential prediction using RNNs. This strategy, called *Scheduled Sampling*, consists in making a transition from a fully-guided scheme, where ground-truths are used as previous tokens, towards a less guided scheme, which rather uses generated tokens. This forces the model to deal with its own mistakes, as it would have to during inference.

In contrast to models with sequentially-generated frames, other authors attempted to yield multiple future frames simultaneously. On the other hand, advances in the field of natural language processing have been a continuous source of inspiration for the developing of these models. For instance, Sutskever et al. [148] introduced a sequence to sequence (*seq-to-seq*) model for machine translation, which was implemented with a encoder-decoder configuration and recurrent units. In addition, it enabled the generation of multiples outputs in parallel. In this groundbreaking work, the encoder is composed by a multilayered Long Short-Term Memory (LSTM) that maps the input sequence to an internal representation called context vector. Subsequently, another LSTM (decoder) is used to generate the output sequence from the vector. The lengths of the input and output sequences can be different, as there is no explicit one-to-one relation between the input and output sequences. Such idea was further extended by [101], who developed a LSTM encoder-decoder framework for image reconstruction and future frame prediction. Luo et al. [9] used a similar configuration for motion prediction. They demonstrate the predictive capability when increasing the horizon. Additionally, they showed how the information contained in the context vector can be used for downstream tasks, such as action recognition (see Figure 3.5).

Other works employ 3D convolutions as the backbone for the encoder and decoder. Mathieu et al. [145] combined all the past frames and feed them into a CNN, thereby finding a mapping function between inputs and future frames. Furthermore, they used an adversarial loss to differentiate fake and real future frames. In contrast with the discriminator of the generative adversarial networks [149], which operates on a single image, this discriminator acts on multiple frames. This introduces an implicit constraint to ensure a temporal consistency. Wu et al. [146] proposed a convolutional autoencoder to generate future frames simultaneously. The authors also used a multi-frame discriminator to ensure temporal coherence. In addition, they integrate an additional discriminator to improve the spatial quality. In summary, parallel models may benefit from parallel computing. Also, for some applications it might be desirable to get multiples frames in one-shot.

### 3.2.4 Pixel synthesis vs. spatial transformations

The approach to generate future frames is a distinctive feature of predictive models. Some models use the decoder to directly synthesize the pixels of future frames [131, 135, 143, 150–

155]. There have been early attempts to sequentially predict RGB pixels by modelling the discrete probability of the raw pixel values and encoding the dependencies in the image [156]. Inspired by this work, Kalchbrenner et al. [157] proposed the Video Pixel Network (VPN) model, which estimated the discrete joint distribution of the raw pixel values in a video.

Generally, pixel-wise regression models assume that the feature representations learned by the encoder contain useful information to reconstruct the future frames. This information is typically passed through skip connections at different scales to ensure good visual quality. Therefore, the powerful U-net architecture [158] is a popular choice for this purpose. The model presented by Castrejon et al. [10] employed a CNN to encode prior frames individually. Subsequently, the extracted features were used to initialize the states of recurrent networks, which were used for decoding purposes. The model design was based on a U-net backbone, where the encoder's features at multiples scales were skipped to the decoder to regress the pixels. Nonetheless, because images are high dimensional and highly structured, direct pixel generation is extremely challenging.

Alternatively, other approaches use spatial transformations on previous frames to yield the future ones [119, 159, 160], thereby avoiding the pixel synthesis. Hence, they leverage the visual appearance already available in the input sequence. The spatial transformation layer in [161] is the essence of vector-based resampling approaches for video prediction. It is fully differentiable and can be integrated at any part of the models [104].

Inpired on this module, Liu et al. [113] proposed the Deep Voxel Flow model. It consists of a multi-scale flow-based encoder-decoder model. Although it was conceived originally for video frame interpolation, the model was also evaluated on a predictive basis reporting sharp results. Similarly, Liang et al. [162] used a warping layer based on bilinear interpolation to warp the last input frame with the predicted flow. Finn et al. introduced the Spatial Transformer Predictor motion-based model [163], which generated 2D affine transformations for bilinear sampling. Furthermore, this work also presented other two kernel-based variants. The first, denoted as Dynamic Neural Advection (DNA), outputs a distribution over locations in the previous frame for each pixel in the new frame. The computed distribution is then used to determine the pixel values. Similarly, the convolutional version of this approach (CDNA) find the parameters of multiple convolution kernels. Then, the next frame is obtained by convoluting the previous image with the kernel.

Instead of focusing on the whole image, Chen et al. [159] followed a local motion modelling approach, i.e. object-centered representations. After selecting the target object, transformation kernels were yielded dynamically as in the DFN [164] and applied to the last patch containing the object. The main limitation of object-centric predictions is that performance

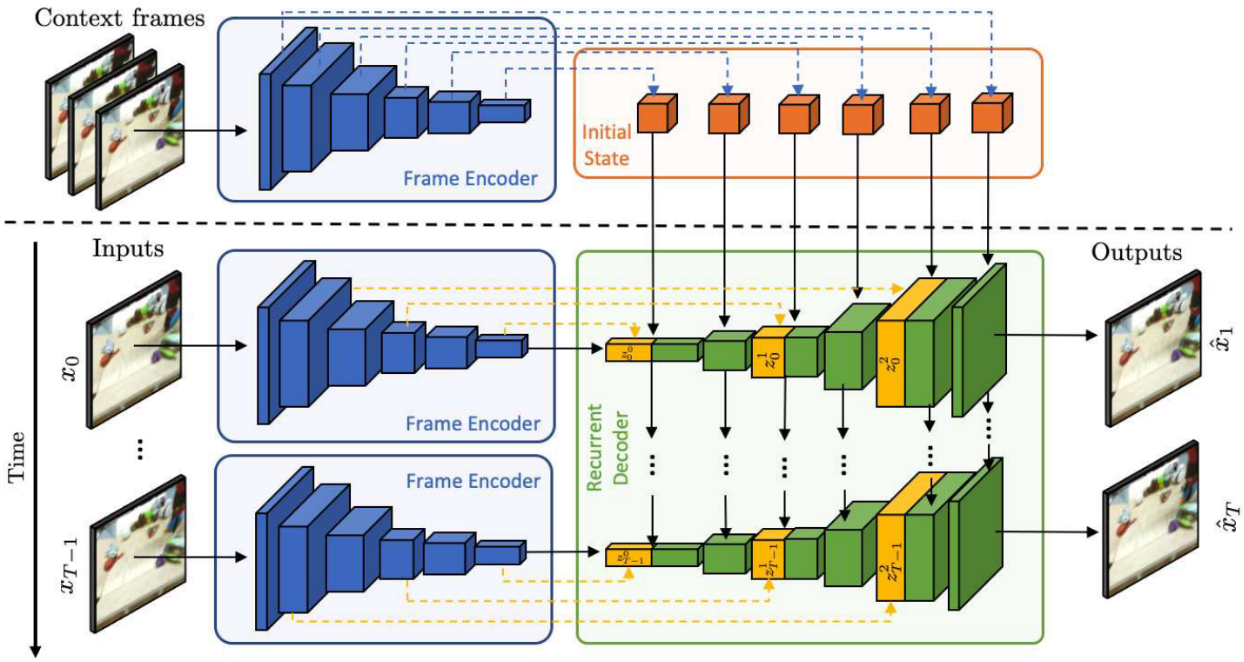


Figure 3.6 Conditional Variational Recurrent Neural Network for video prediction. Source [10]

decrease when dealing with multiple objects and occlusions [104]. Overall, transformation-based models lead to sharper results. Their main limitation is the incapacity to generate structures that are not contained in the previous (source) images.

### 3.2.5 Deterministic vs. stochastic models

In deterministic models, the output is fully determined by the parameter values and the initial conditions. This means that the model will always provide the same results for a given input. Modeling the randomness of future events is extremely challenging. Hence, most of the video prediction models are deterministic [132,133,146,151]. For instance, PredNet [165] and ContextVP [166] are deterministic models, based on recurrent architectures, that have been proposed for video prediction of car mounted scenes.

In contrast, in stochastic models a certain set of parameters and initial conditions will yield an ensemble of different outputs. Thus, randomness is inherent to them. An early work in stochastic modeling was developed by Babaeizadeh et al. [130]. These authors proposed a stochastic variational video prediction (SV2P) method, which is able to provide a different prediction for each sample of its latent variables. At training time, the inference network estimates the posterior using both past and future images. The latent value $z$, sampled from the posterior distribution, is fed to the generative network. Optionally the action can be also passed. The generative network, taken from [163], predicts the next frame given the previous frames, latent values, and actions. At test time, $z$ is sampled from the prior.

Further works continued developing the idea of leveraging future frames as prior knowledge during training [144, 167]. Lee et al. [167] combined the advantages offered by both GAN and VAE within an stochastic adversarial video prediction (SAVP) model. Specifically, they added a GAN architecture into the aforementioned model (SV2P) [130] to improve the visual quality of future frames. Denton et al. [144] combined a deterministic frame predictor with time-dependent stochastic latent variables. They developed an inference network, based on LSTM, to estimate the latent distribution for each time step recursively. During training, they leveraged the ground-truth images to compute a prior over the latent variables, which can be fixed or learnable. The fixed prior is generally assumed to follow a Gaussian distribution while the other alternative learns the prior for the latent variable $z$ from the input frames. These prior-based models showed that using the future frames as additional information during training can boost the results. The authors argued that the learned prior can be interpreted as a predictive model of uncertainty.

Indeed, several works have attempted to deal with uncertainty, either in the inputs [134] and outputs [137]. The work presented by Jang et al. [137] tried to answer the question on how should a model behave when there are multiple correct, equally probable future images. They propose a generative adversarial network conditioned on appearance and motion information. The model is composed by generator, two discriminators for the appearance and motion pathways, and a perceptual ranking module that encourages videos of similar conditions to look similar. Wang et al. [134] argued that previous works assume the spatiotemporal coherence of the inputs and fail to deal with perceptual uncertainty. In other words, they do not work well for noisy inputs, where spatiotemporal consistency is significantly broken. Therefore, they proposed a Bayesian Predictive Network (BP-Net), which is able to cope with both perceptual and dynamic uncertainties. The model combines Bayesian inference and recurrent neural networks. Similarly to [144], the predictions are aided by a learned prior. Other probabilistic approaches have also been proposed to cope with uncertainty [130, 167, 168].

### 3.3 Respiratory motion models

Motion models establish a relationship between some surrogate data (input) and a motion estimate (output) [47]. It often finds an application when measuring the motion of interest is not feasible. For instance, due to limitations in the temporal resolution during certain procedures. When a motion model is used, measurements are made of some surrogate data, also known as partial observations, instead of measuring the motion of interest directly. Such a signal is required to have a strong correlation with the motion of interest and be easily measurable. There are different sources of surrogate data. Some of them were mentioned in Section 3.1.1. In the literature, several motion models have been introduced which can be divided into subject-specific and population-based models, depending on the data used for their creation.

### 3.3.1 Subject-specific motion models

In a early work, Blackall et al. [31] constructed a subject-specific statistical model of the liver by using non-rigid registration. The resulting free-form transformations were used to propagate landmarks derived from a segmented liver surface of a template image to the other images throughout the breathing cycle. Principal component analysis of these landmarks was used to produce a statistical model of motion and deformation. The maximum deformation captured by the model was approximately 15 mm for deep breathing and 10 mm for shallow breathing. In [169] the same authors built a subject-specific motion model to constrain the alignment between 3D preoperative images (either MR or CT) and intra-operative US during thermal ablation of liver metastases. To this end, six breath-hold volumes were acquired at different phases in the respiratory cycle. A selected reference volume was registered to other volumes across the respiratory cycle using free-form deformation based on B-splines. The six images were assigned equally spaced T-values starting at $T = 0.0$ (end-exhale) and $T = 1.0$ (inhale). Subsequently, third-order polynomials in $T$ were fitted to the 3D motion vectors and the coefficients for each of these polynomials were recorded. This formed a model that allowed to interpolate the deformation at any time between end exhale and inhale given the value of $T$.

Similarly, Rohlfing et al. [11] proposed to quantify abdominal organ deformations using intensity-based nonrigid image registration. They introduced the idea of applying the temporal sequence of deformations to a geometrical model to determine the position and shape of a reference image throughout the respiratory cycle (see Figure 3.7). They experimented with MR liver images from four healthy volunteers. Although this method is not feasible for real-time treatment, it was an important step towards the development of motion models.

Zhang et al. [170] employed a deformable registration algorithm to align respiration-correlated CT volumes to a reference volume from four lung cancer patients. An additional input of their method was the diaphragm positions at ten phases of the respiratory cycle. A principal component analysis was performed to parameterize the 3D deformation field in terms of the

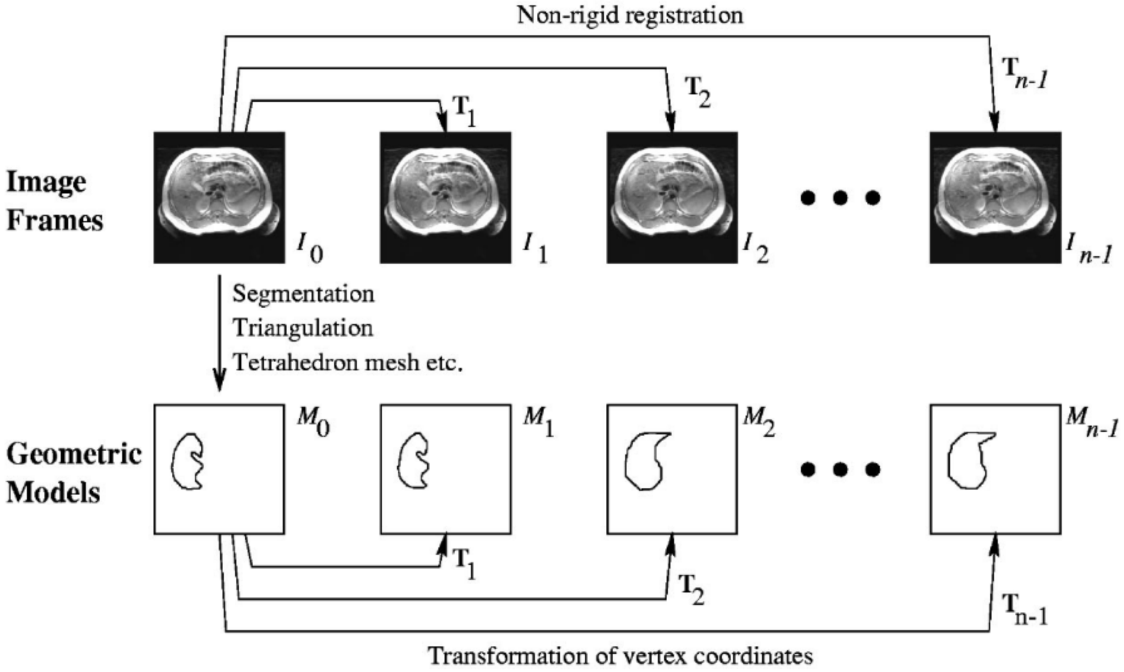


Figure 3.7 Non-rigid image registration and model deformation. All frames are registered to a common reference $I_0$. A geometrical model $M_0$ is generated from the reference image and deformed using the transformations determined by intensity-based registration. The model could be the external skin surface, liver surface, or an internal liver structure such as a radiosurgical target (e.g., tumor). Source [11]

diaphragm motion. They showed that images artifacts, that commonly occur at the mid-respiration states, were reduced in the model-generated images. This approach may have limitations in cases where the correlation between lung tumor and diaphragm position is less reliable such as superiorly located tumors and inter-fraction changes in tumor-diaphragm correlation.

Eom et al. [171] proposed a nonlinear finite element model of respiratory motion during a full breathing cycle based on patient specific pressure-volume relationship and 4D CT data. To achieve a physiologically plausible respiratory motion modeling they used thee pressure-volume (PV) relationship to apply pressure loading on the surface of the model. An experimental hyperelastic soft tissue model was used. The validation was performed using 51 landmarks from CT data. The average differences in position were 0.07 cm (0.20 cm), 0.07 cm (0.15 cm), and 0.22 cm (0.18 cm) in the left-right, anterior-posterior, and superior-inferior directions, respectively.

King et al. [12] built a subject-specific model aimed at improving the acquisition of PET images during free-breathing. For the motion model creation they register dynamic volumes

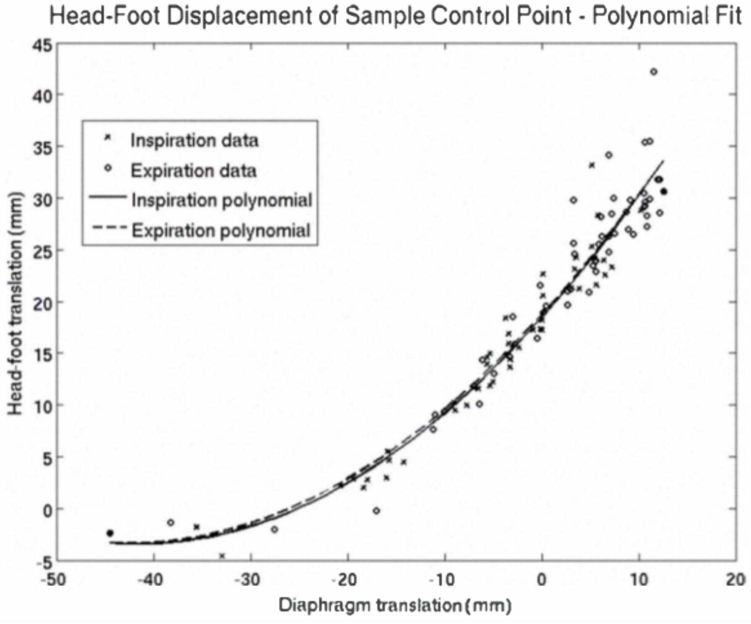


Figure 3.8 Plot of constrained inspiration and expiration polynomials for the head-foot translation of a sample control point. There is little hysteresis as the inspiration and expiration polynomials are very similar. The full deformation motion model consists of 3 pairs of polynomials such as these (for $x$, $y$ and $z$ displacements) for each control point. Source [12]

to a reference end-exhale volume by using hierarchical local affine registration. The motion fields were modelled as second order polynomial functions of a 1D surrogate signal using a least squares technique. In this case, the used surrogate was the head-foot diaphragm translation. In this way, given a value for the respiratory surrogate and a breathing direction, the model estimates the 3D displacements vectors. Figure 3.8 illustrate the motion model.

Noorda et al. [172] developed a subject-specific model of one average motion cycle of the entire liver. They registered dynamic MR slices of six anatomical positions to a 3D volume. The obtained deformation fields were clustered according to their respiratory phase. They were then averaged for every location and interpolated in 3D, to yield a 3D deformation field for every cluster. The liver was then deformed according to these 3D deformation fields, to obtain a look-up table of the liver for all possible states. The average error in the prediction of the blood vessel center positions was 3.0 mm.

Broadly speaking, statistical modelling is the most commonly used technique. Recent works have attempted to relate image surrogates to the model coefficients [90, 173–175]. Generally, these works rely on the maximization of a similarity metric between the image and the corresponding slice position taken from a reference volume, which is iteratively warped until convergence is reached.

Principal Component Analysis is a well-known dimensionality reduction technique, which is useful for exploratory data analysis and for predictive modelling. It projects the data points onto the first few principal components to obtain lower-dimensional data while preserving the variation of the data. Stemkens et al. [90] applied PCA to parameterize 3D deformations. The weights of the eigenvectors were iteratively optimized until achieving the best alignment between a warped reference volume and the surrogate slices. Similarly, [175] proposed to refine the motion model using free-form deformations.

Such an approach has been extensively validated using MRI and kV projections [173–177]. Following the same approach, [178] created a PCA model to establish correlations between 2D navigator images and 3D displacements. Further, Garau et al. [179] proposed a variant of this work by using a region-based approach for better local adaptation. In an early attempt to leverage deep neural networks in the context of respiratory motion modelling, Giger et al. [180] presented a conditional generative adversarial network to relate an ultrasound image with a future deformation. This work was only validated to work in subject-specific condition.

In the literature, tracking errors reported for patient-specific models are often lower than those for population models. Nevertheless, there are some limitations for their use in the clinical context, as its reliability depends on an accurate response to inter-fraction motion variations. Furthermore, due to time constraints, in many clinical scenarios it is not possible to acquire a patient-specific 4D dataset and non-rigidly register the volumes to create the model just before treatment.

### 3.3.2 Population-based motion models

To build statistical population models, a crucial step is the establishment of inter-subject correspondences. Specifically for motion analysis, it is important that corresponding points undergo the same breathing-induced deformation. Therefore, a mechanical correspondence needs to be established. There are multiples strategies to establish such a correspondence. According to [181], manual labeling, distance-based correspondence schemes, which favour the correspondence of close points after a certain alignment such as Procrustes matching, or correspondence based on parameterization are the most used techniques.

Von Siebenthal et al. [13] captured the deformation of the liver at exhalation from 12 volunteers through non-rigid registration. With this data, they created a statistical motion model. First, the liver was segmented manually in one exhalation volume per subject. This yielded fine triangular surface meshes with several thousands of triangles. Four landmarks were manually labelled in each sagittal slice as shown in Figure 3.9 and connected by B-splines. Subsequently, a coarse mesh prototype of the right liver lobe was then aligned to the fine

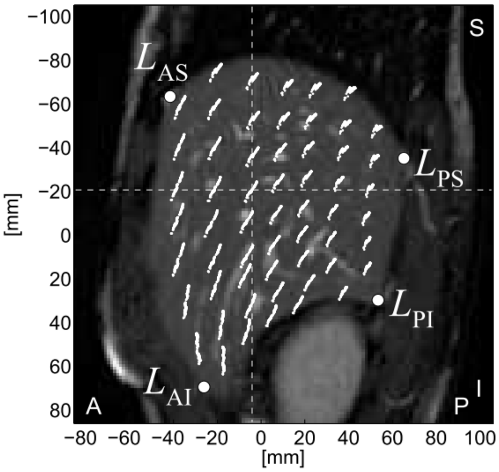


Figure 3.9 Four landmarks $L_{AI}$, $L_{AS}$, $L_{PI}$, $L_{PS}$, where the indices indicate the location (anterior, posterior, superior, inferior) were manually placed. These points mark the delineations between the superior surface in contact with lung, the anterior and the posterior areas, which slide along the abdominal wall, and the inferior surface. Source [13]

mesh of each specific liver such that its four edges coincided with the marked delineations (see Figure 3.10(a)). The vertices of the prototype were regularly distributed along the landmark splines in the medio-lateral direction. The coarse prototype was then gradually refined to fit the fine surface mesh (Figure 3.10(b)). Further refinement steps were performed for each of the 12 livers (Figure 3.10(c)). A regular grid of 290 points with 15 mm resolution was placed in the average liver shape and then transformed to each subject-specfic liver. From this data, a point distribution model was built. Specifically, PCA was applied to determine the eigenvectors of the covariance matrix. It is important to emphasize that this model only captured the deformation in exhale positions. Results showed that there are three typical modes of deformation during treatment and the maximum displacement was 5 mm or larger in all subjects and ranged up to 18.8 mm.

Arnold et al. [182] presented an atlas-based prediction technique built using 4D MRI data. It was combined with the population-based statistical exhalation drift model explained previously ( [13]) to account for the non-periodic slower organ drifts. This approach was able to capture the full patient specific motion of the organ. Based on a breathing signal, the respiratory state of the organ is then tracked and used to predict the target's future position. The method was validated on the same dataset used in [13]. The prediction of the liver positions resulted in an average error of 1.1 mm over time intervals of up to 13 minutes.

Samei et al. [33] proposed the use of exemplar models as a non-parametric method for adapting the population model to an individual subject during therapy. They gathered 4D MRI

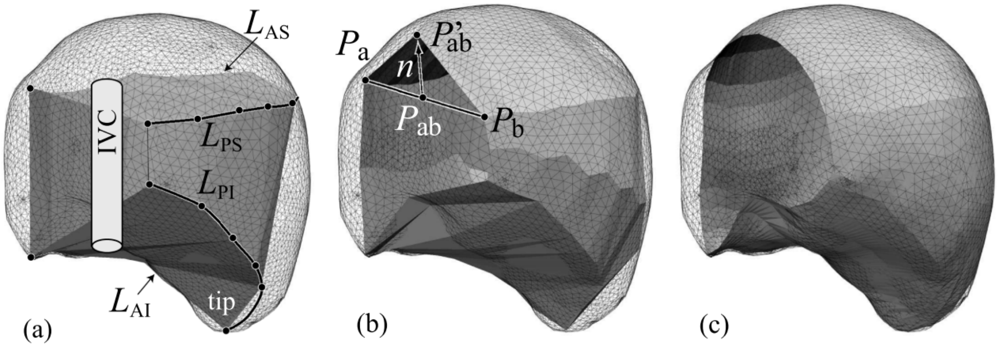


Figure 3.10 (a) Coarse mesh prototype aligned with respect to manually identified delineations on the liver surface. (b) Refined mesh after correspondence preserving subdivision and projection. (c) Resulting mesh after three refinement steps. Source [13]

data from 12 volunteers and quantified the liver motion using intensity-based non-rigid registration. Inter-subject correspondences were established following the same approach of the aforementioned work [13]. They built exemplar models by fitting a PCA model to the motion vectors of each individual subject. The final model was a weighted combination of the predictions of all the sub-models. The weights were based on the squared Mahalanobis distance between the surrogate and the corresponding components of an individual model. They found that the use of exemplar models improves the lowest error achieved by the PCA model by 10%. This liver motion model had subsequently 2 follow-up works. First, Tanner et al. [183] improved the distance measure by taking into account the history of the surrogates. Also they proposed the individualization of the exemplar model by a subject-specific example 3D motion field, which was extracted from an additional end-inhalation image. In a second work [184], the same model was validated by using a 3D breath-hold image and an interleaved acquisition of two MR slices. From those two slices, one was used for tracking and the other for validation of the prediction accuracy. The motion of the liver on the validation slice was spatio-temporally predicted with an accuracy of 1.9 (4.4) mm for a latency of 216 ms.

Preiswerk et al. [32] used the same dataset [13] to predict liver motion by using a statistical population model. They also found mechanical correspondences between subjects with four manually annotated landmarks in the same positions as illustrated in Figure 3.9. They neglected the left liver lobe because it is heavily influenced by the motion of the heart, which was not in the scope of their study. For each subject, they obtained a vector of corresponding surface points at exhalation. Then they aligned those vectors using partial (no scaling) generalized Procrustes analysis into a common coordinate system. Their model was based on PCA and achieve an average prediction error of 1.2 mm.

In contrast to all the aforementioned works, some authors have explored biomechanical models. Brock et al [185] employed Finite element Analysis (FEA) on two liver CT scans to construct a 4D model of the liver during breathing. A linear elastic, small deformation mechanical model was applied to one patient to obtain an intermediate organ position and shape between exhale and inhale. Known transformations between anatomically defined subsections of the exhale and inhale liver surfaces were applied as constraints to the exhale CT liver model. Intermediate states were then calculated and time weighted to determine a 4D respiratory model of the liver.

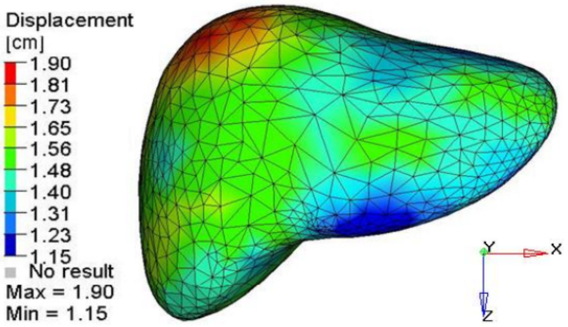


Figure 3.11 A colour scale of the population deformation map applied to the exhale population finite element liver model to generate a liver population respiratory motion model. Source [14]

Following a similar approach, Nguyen et al. [14] used 10 patients exhale and inhale breath-hold images and meshes to generate a population motion model. They performed rigid registrations to align exhale meshes to one arbitrary chosen mesh. The transformations were applied to binary masks and all the contours were summed and converted into a 3D triangular finite element surface mesh as depicted in Figure 3.11. The population liver model was deformed to match each patient's exhale and inhale using the biomechanical framework MORFEUS [186]. This resulted in a deformation map describing the patient's specific respiration motion using a common set of elements. An average deformation map was then calculated as the mean deformation at each node. The average respiratory deformation map was then applied to the population exhale liver model to generate a liver population respiratory motion model.

Other approaches such as kernel PCA, support vector machine, atlas and Kalman filtering have been involved in the creation of related models. He et al. [187] generated a 4D motion model of the lung from dynamic CT volumes of 30 patients. First, they segmented the lungs and aligned the surfaces. Kernel PCA (K-PCA) was then applied on the lung field motion vectors derived from the extracted lung field surfaces to model the motion. They trained a support vector machine (SVM) to model the relation between motion of fiducial markers and the coefficients of the K-PCA, which contain the representation of lung's surface motion. During the intervention, the trained SVM motion model is used to estimate the lung motion vectors from real-time fiducial signals. These motion patterns can be used to estimate the patient-specific serial CTs from a static 3D CT and the real-time respiratory signals of that patient. The reported average accuracy was 1.63 mm.

Ehrhardt et al. [188] modeled the lung motion from thoracic 4D CT data of 17 patients. The process consisted of three steps: an intra-subject registration to generate subject-specific motion models, the generation of an average shape and intensity atlas of the lung as anatomical reference frame, and the registration of the subject-specific motion models to the atlas in order to build a statistical 4D mean motion model. The prediction was evaluated with respect to landmark and tumor motion. The mean target registration error (TRE) was $3.3 \pm 1.6$ mm in cases where lung dynamics are not impaired by large lung tumors or other disorders.

Ries et al. [189] proposed a real-time tracking method that observes the target on a 2D image plane combined with a perpendicular pencil beam navigator, finally obtaining 3D information of the targets trajectories. The future target position is then estimated by a 3D Kalman filter. The method was tested in phantom experiments on human kidneys and in-vivo with kidneys of ventilated pigs, both following a regular and stable breathing pattern.

## 3.4 Summary

In this chapter, we conducted a literature review of different topics, which are relevant for the research project. First, we presented an overview of current methods for 4D imaging, which can be classified in two main groups according to the type of readout. In this review, we have observed that multi-slice 2D acquisitions have been widely explored while interest in 3D acquisitions is increasing rapidly. In the former case, one of the main challenges is the acquisition of navigator signals, which are crucial for the slice sorting. Some current self-sorting methods still involve human intervention. For instance, for the manual identification of a reference image at each anatomical position, which is unfeasible for large datasets. Another shortcoming is the requirement of an organ segmentation.

The second topic was about image-based temporal prediction methods. We described different strategies for future image generation along with their limitations. This task is not exempt from hurdles such as the prediction from limited dynamics, and the high-dimensionality inherent to complex deformations. Moreover, while extensive research have been done for natural images, contributions suited to the challenging medical datasets are more scarce.

Finally, we presented the foundation of respiratory motion models as well as their categorization according to the data involved in their creation. In general, subject-specific models are often more accurate than the population-based. However, the assumption of a pre-treatment 4D dataset acquisition and processing constitutes an important limitation since it is not always available in many clinical scenarios. In contrast, population-based models can be applied in new/unseen patients in the absence of any 4D image data. Nevertheless, the establishment of inter-subject correspondences remains the main obstacle.

# CHAPTER 4 RESEARCH METHODOLOGY

## 4.1 Problem statement

Analyzing breathing-induced organ deformation requires, first of all, acquiring ground truth motion data. Based on the observations, the next step is to create a motion model and relate it to surrogate signals. Finally, a mechanism for temporal forecasting, acting on the surrogates, is required to cope with system latencies.

The basis for building respiratory motion models consists of imaging and estimating ground-truth motion data, which shows the moving organs during free-breathing. This means that 3D+$t$, also known as 4D, data is needed. However, the current technology has limited capacity to fulfill this criteria due to spatiotemporal constrains. Retrospectively sorting of slice-base acquisitions is a solution which requires determining the respiratory state of each slice using navigator signals. External navigators may not perfectly correlate with the actual motion while the internal ones increase the required temporal resolution and prolong the acquisition time. Existing self-sorting methods often require manually defining a reference image across the acquired anatomical positions, which is time-consuming and requires trained experts. Hence, the first research question is formulated as: **Is it possible to design a fully automatic 4D volume reconstruction self-sorting technique, which is able to preserve the temporal consistency and capture the free-breathing organ motion?**

After estimating the motion in a population of free-breathing subjects, the next step is to construct the motion model by finding the principal modes of variations in a low-dimensional space to better understand the underlying process. The main challenge in creating population-based motion models is the establishment of inter-subject correspondences, which typically involves the organ segmentation, shape meshing and finding mechanical corresponding landmarks across subjects. This step often requires manual intervention and can be complex and time-consuming, especially in large datasets. Therefore, the following questions arises: **Could the strong generalization capabilities of deep neural networks replace this complex step? If so, what architecture would be suitable for motion modelling and how surrogates signals would be related to a personalized patient model?**

The time required for image acquisition, target localization, and subsequent beam modulation/tracking adds a significant cumulative latency to the system. In consequence, during real-time treatments, by the time a gating decision has been made the patient anatomy has already changed. Generally, temporal forecasting is integrated within motion models to compensate these latencies. Therefore, the final question is: **Is it possible to design an image-based temporal predictor, able to yield ahead-of-time visual representations corresponding to future organ states, that can be used by the motion model to generate future volumetric deformations?**

## 4.2 Hypothesis

Considering the problems stated in the previous section, the following hypothesis can be formulated.

**Hypothesis 1:** An automatic slice reordering methodology, comprising the extraction of a pseudo-navigator from the cine-acquisitions and the subsequent slice stacking, can be designed to construct 4D volumes from navigator-less MR images.

**Hypothesis 2:** A motion model, constructed from deep neural networks, can be created to learn typical patterns over a population and then generalize in unseen anatomies.

**Hypothesis 3:** An image-based spatiotemporal predictive mechanism, based on deep neural networks, can be developed and integrated to the motion model to generate volumes corresponding to future times.

## 4.3 Objectives

The general objective of this thesis is the **development of a 4D motion modelling framework using deep learning allowing the generation of temporal volumes from surrogate 2D images** in the context of image-guided radiotherapy. The main goal is accomplished through the following specific objectives.

**Objective 1:** Developing an automatic slice reordering methodology to construct 4D volumes from navigator-less MR images.

**Objective 2:** Designing motion modelling solutions, relying on deep neural networks, able to learn over a population dataset and generalize in unseen anatomies.

**Objective 3:** Designing an image-based temporal predictor and integrating it within the motion model to enable future volume generation.

## 4.4 General methodology

In this research, an ensemble of novel solutions are described for the analysis and modelling of respiratory organ motion. Figure 4.1 depicts an overview of the entire proposed framework, which consists of three main blocks: (1) Automatic volume construction from navigator-less cine acquisition (2) Motion modelling framework; and (3) Image-based temporal prediction.

The methodological structure of this thesis is illustrated in Figure 4.2. The first objective of this project refers to the design of 4D imaging strategies. Specifically, to the design of an automatic slice reordering methodology to construct 4D volumes from navigator-less cine acquisitions. This methodology, presented in Chapter 5 (Romaguera et al. [80]), introduces an approach to extract pseudonavigator signals from cine acquisitions. In addition, it presents a graph-based method for slice stacking. In a related work, which is presented in Appendix A (Romaguera et al. [74]) another slice reordering technique is designed by combining image similarity measures with manifold alignment theory. In this method, the slice corresponding to each point in the low-dimensional manifold is compared to the slices corresponding to its neighborhood. These approaches allow us to obtain temporal volumes from navigator-less cine acquisitions with no human intervention. The estimated deformations between a fixed volume and temporal volumes in the 4D dataset constitute the basis for motion model creation.

The second objective is to design data-driven motion modelling solutions by learning from population data. The proposed motion models are based exclusively on deep neural networks and do not require any prior steps such as manual segmentation or landmark identification. During their creation, they perform dimensionality reduction on the input 3D+$t$ deformations. Furthermore, they establish an approach to relate partial observations with the predictions. A first deterministic variant, presented in Chapter 6, relies on convolutional autoencoding as a backbone for the modeling task (Romaguera et al. [190]). Volumetric organ deformations and surrogate slices are mapped to a common latent space, where they both are associated by minimizing their point-wise distances. A probabilistic version of the motion modelling task is described in Chapter 7, which is formulated as a conditional manifold learning task (Romaguera et al. [191]). Specifically, this work propose the integration of feature vectors, extracted from the pre-operative volume, and visual representations, extracted from the surrogate images, as predictive variables to predict future volumetric deformations. The personalization capability of this probabilistic model was explored in another work presented in Appendix B (Romaguera et al. [192]). These motion models are tested on datasets acquired both in healthy volunteers and cancer patients using different imaging modalities. In contrast to classical approaches, they do not requires establishing inter-subject correspondences.

Hence, the burden of this complex step is removed and is rather replaced by unsupervised feature learning across population samples.

The third objective is to investigate possible structures for temporal prediction from an input image sequence. Specifically, the predictive mechanism should be able to forecast

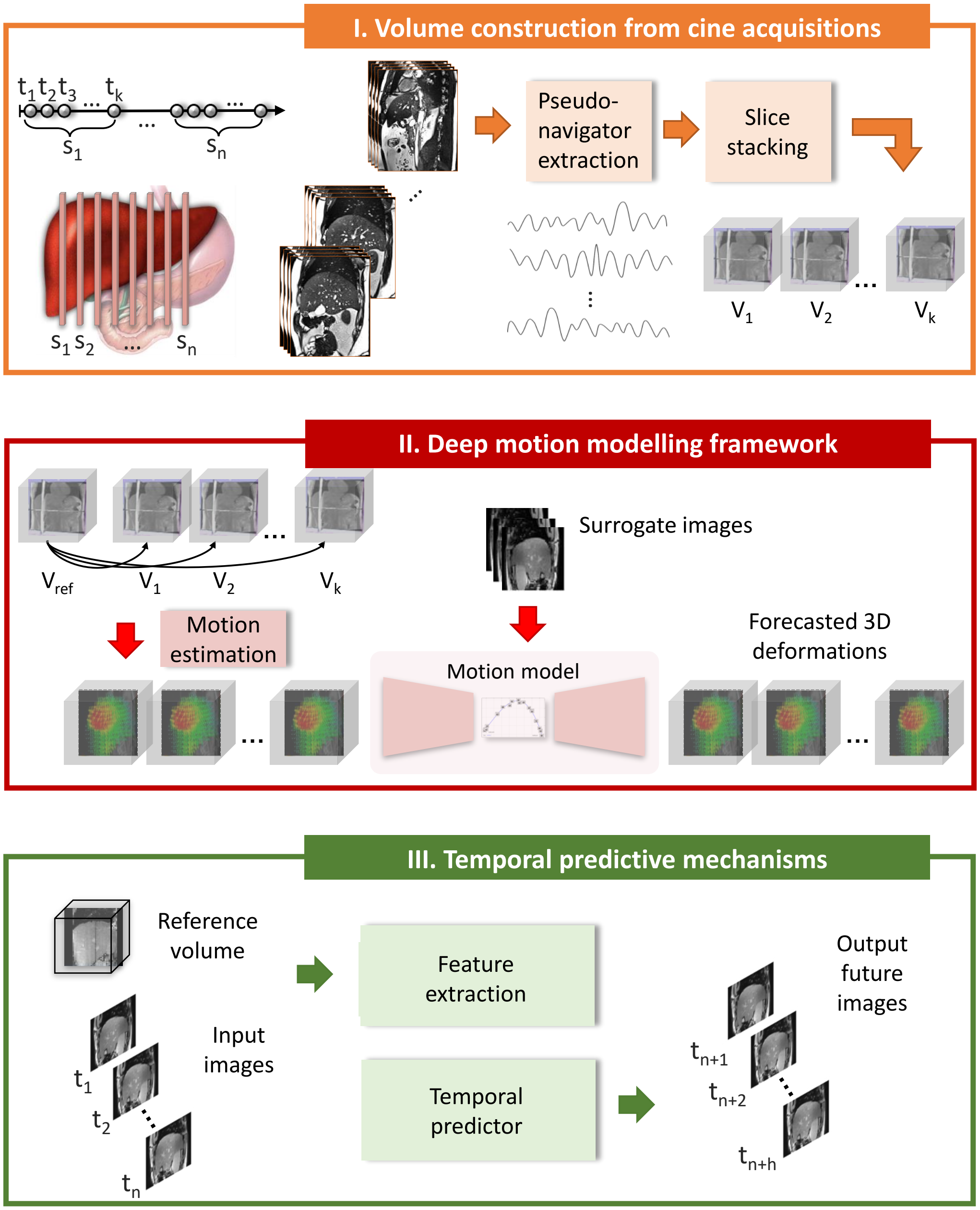
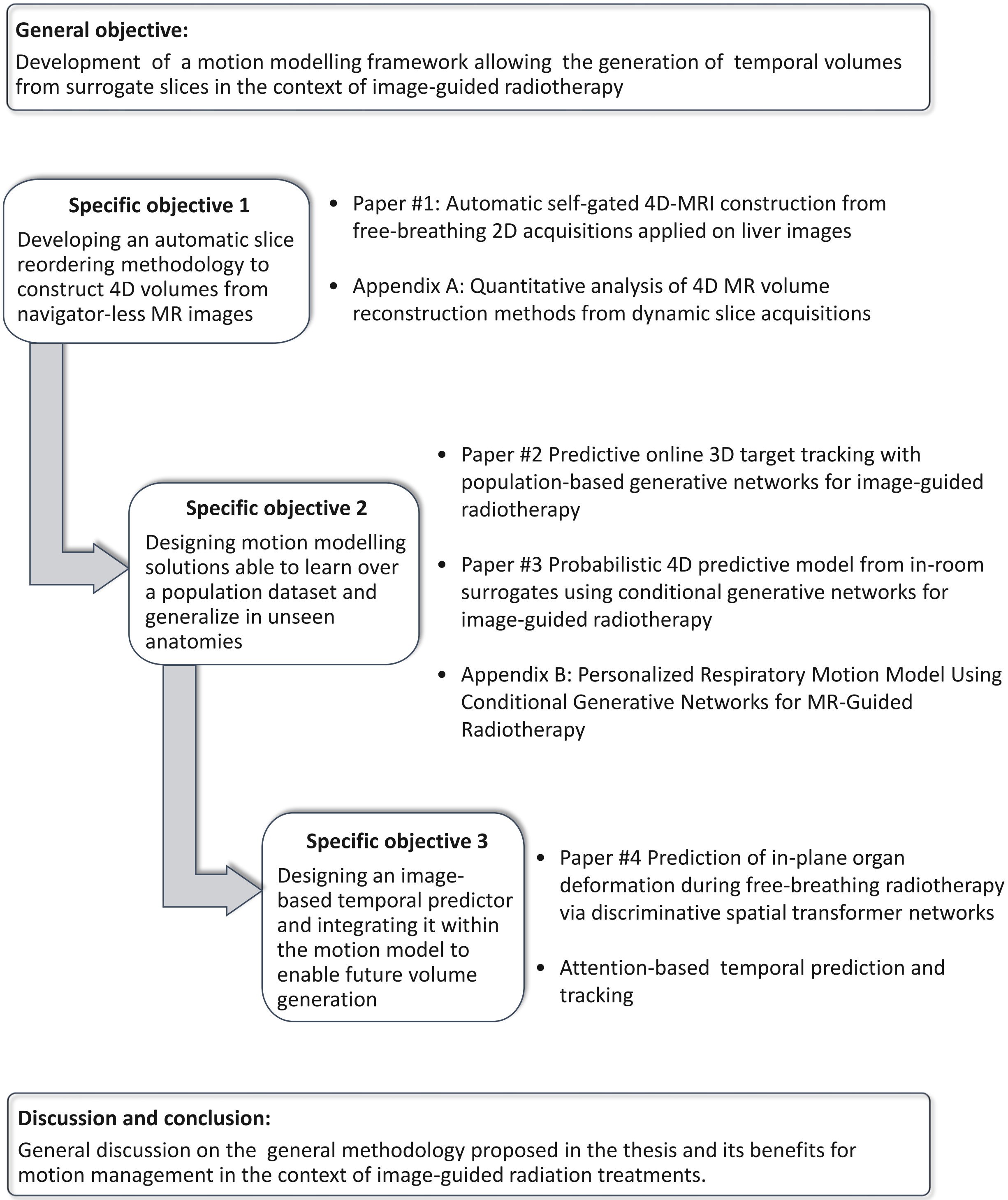


Figure 4.1 Proposed framework. (1) Volume reconstruction: The inputs are cine-acquisitions, and the output is a 3D+t dataset; (2) Motion modelling: The input is ground-truth motion data, and the output is a reconstructed version of the input; (3) Temporal prediction: The input is an image sequence, and the output is an image sequence at future times.

visual representations from the spatiotemporal information contained in the dynamic 2D images, that can be recovered as future deformations. The model presented in Chapter 8 (Romaguera et al. [119]), relies on multi-scale feature extraction, convolutional recurrent units and spatial transformations to implicit regress future images. Another work, exposed in Chapter 9, explores an attention-based model to predict future representations from an image sequence. Moreover, it integrates prior knowledge from future frames available during model training to regularize the latent representations. Additionally, this work demonstrates how the previously created model can be leveraged for motion-compensated 3D target tracking.

The methods proposed across the different steps of the project, as well as the obtained results, are presented by means of articles in Chapter 5 (objective 1), Chapters 6 and 7 (objective 2), and Chapters 8 and 9 (objective 3). Finally, the discussion, conclusion, and recommendations are presented in Chapters 10 and 11.



**General objective:**

Development of a motion modelling framework allowing the generation of temporal volumes from surrogate slices in the context of image-guided radiotherapy

**Specific objective 1**

Developing an automatic slice reordering methodology to construct 4D volumes from navigator-less MR images

- Paper #1: Automatic self-gated 4D-MRI construction from free-breathing 2D acquisitions applied on liver images
- Appendix A: Quantitative analysis of 4D MR volume reconstruction methods from dynamic slice acquisitions

**Specific objective 2**

Designing motion modelling solutions able to learn over a population dataset and generalize in unseen anatomies

- Paper #2 Predictive online 3D target tracking with population-based generative networks for image-guided radiotherapy
- Paper #3 Probabilistic 4D predictive model from in-room surrogates using conditional generative networks for image-guided radiotherapy
- Appendix B: Personalized Respiratory Motion Model Using Conditional Generative Networks for MR-Guided Radiotherapy

**Specific objective 3**

Designing an image-based temporal predictor and integrating it within the motion model to enable future volume generation

- Paper #4 Prediction of in-plane organ deformation during free-breathing radiotherapy via discriminative spatial transformer networks
- Attention-based temporal prediction and tracking

**Discussion and conclusion:**

General discussion on the general methodology proposed in the thesis and its benefits for motion management in the context of image-guided radiation treatments.

Figure 4.2 Methodological organization of the thesis.

# CHAPTER 5 ARTICLE 1: AUTOMATIC SELF-GATED 4D-MRI CONSTRUCTION FROM FREE-BREATHING 2D ACQUISITIONS APPLIED ON LIVER IMAGES

Contribution of the first author in preparation and writing this paper is evaluated as 90%. This article has been published by the International Journal of Computer Assisted Radiology and Surgery on March 2019.

**Remarks:** This paper presents an automatic weighted graph-based method designed for volume reconstruction from navigator-less cine acquisitions. The proposed method derives a pseudo navigator signal from temporal slices and creates a weighted graph to guide the slice stacking process. Experiments revealed that the proposed approach is able to automatically detect the end-exhale phases within the temporal slices at one given anatomical position, and cope with irregular breathing during the sorting process.

## Automatic self-gated 4D-MRI construction from free-breathing 2D acquisitions applied on liver images

Liset Vázquez Romaguera[1], Nils Olofsson[2], Rosalie Plantefève[3], Elodie Lugez[4], Jacques De Guise[2], Samuel Kadoury[1,3]

[1] École Polytechnique de Montréal, [2] École de Technologie Supérieure [3] Centre de recherche du Centre Hospitalier de l'Université de Montréal, [4] Elekta Ltd., Montréal, Canada

## 5.1 Abstract

*Purpose* MRI slice reordering is a necessary step when three-dimensional (3D) motion of an anatomical region of interest (ROI) has to be extracted from multiple two-dimensional (2D) dynamic acquisition planes, eg. for the construction of motion models used for image-guided radiotherapy. Existing reordering methods focus on obtaining a spatially coherent reconstructed volume for each time. However, little attention has been paid to the temporal coherence of the reconstructed volumes, which is of primary importance for accurate 3D motion extraction. This paper proposes a fully automatic self-sorting four-dimensional (4D) MR volume construction method that ensures the temporal coherence of the results. *Methods* First, a pseudo navigator signal is extracted for each 2D dynamic slice acquisition series. Then, a weighted graph is created using both spatial and motion information provided by the pseudo navigator. Volume at a given time point is reconstructed following the shortest paths in the graph starting that time point of a reference slice chosen based on its pseudo navigator signal. The proposed method is evaluated against two state of the art slice reordering algorithms on a prospective dataset of 12 volunteers using both spatial and temporal quality metrics. *Results* The automated end exhale extraction showed results closed to the median value of the manual operators. Furthermore, the results of the validation metrics show that the proposed method outperforms state of the art methods in terms of both spatial and temporal quality. *Conclusions* Our approach is able to automatically detect the end-exhale phases within one given anatomical position and cope with irregular breathing.

**Keywords** Slice reordering, 4D image construction, Motion extraction, Liver, MRI

### 5.2 Introduction

Enabling free-breathing liver cancer therapies such as external beam radiotherapy requires accurate tracking of the internal anatomy and tumor location during treatment in order to focus radiation beams to targets and avoid surrounding anatomy. Typically, a radiation oncologist will use multiplanar (in axial and coronal planes) images of the targeted organ and its internal and surrounding structures (blood vessels, kidney) acquired prior to intervention or a few images during the procedure. However, a major limitation of vascular and focal interventions resides in the patient's respiration or involuntary movement, which may stray the pre-defined target and trajectories determined during planning from the actual anatomy, thus inducing errors in the relative position of the instrument performing the action with respect to the target. Furthermore, live motion tracking of the internal anatomy depends on 3D imaging and image post-processing in real-time, which is unfeasible during interventional procedures. Thus, to complete partial information (2D images, navigator signal) clinically available during treatment, prior knowledge of the anticipated motion field during the breathing cycle is necessary [193]. Some intraoperative acquisitions require contrast agent, which is problematic as it will increase the toxicity to the patient. Clinicians therefore avoid using these intraoperative images as much as possible: they visually measure how the tumor target moves with few intraoperative images at the beginning of the intervention, and then use their intuition of the internal motion of tumor with regards to the therapeutic tool in order to achieve a proper targeting. Consequently, there is a clear clinical need to accurately track tumor displacement during free-breathing radiotherapy.

Imaging volumes over time is not a feasible option since it compromises spatial and temporal resolutions [79]. Therefore, most of the four-dimensional (4D) imaging approaches are based on retrospective sorting of computerized tomography (CT) or magnetic resonance (MR) dynamic two-dimensional (2D) slice series according to their respiratory phases. MR imaging presents two main advantages with respect to CT: it does not emit ionizing radiation and offers a higher soft tissue contrast. The latter is crucial as some tumors surrounded by soft tissue with similar density may not be well visualized in CT [194].

Intra-operative respiratory phase tracking is typically achieved using external and internal (anatomical landmarks) surrogates [195]. However, these methods are subject to numerous drawbacks. For instance, external respiratory devices like respiratory belts are known to have a low correlation with the internal organ motion, causing various artifacts in the resulting sorted 4D images. Moreover, internal surrogates, such as the 1D navigator echo, decrease the temporal resolution and may cause interference during the acquisition process. To alleviate these shortcomings, a motion signal using features contained in the captured images can be used. This is known as self-sorting or self-gating methods.

Currently, the acquisition of radial k-space is gaining attention [196–198]. One of its main advantages is its relative insensitivity to motion artifacts at the cost of signal to noise ratio. Although this technique seems promising, it is still in its infancy and needs further research. In addition, pulse sequences will have to be approved for clinical use before they can be used routinely on patients.

Slice reordering techniques that do not rely on external or internal surrogate signals can be grouped into two main categories: machine learning [5, 69–73] and slice feature extraction-based methods [6, 68, 75–79, 199]. The work presented in [200] combine radial k-space acquisition and manifold learning.

Manifold learning (ML) techniques have been employed to map acquired data in a low-dimensional space according to their respiratory phases. This approach commonly makes assumptions about the regularity of the respiratory motion. However, there is a non-negligible residual variability that makes data discrimination difficult. This limitation is common to all cited methods [5, 69–73], which makes it difficult to handle outliers or extreme peaks in the respiratory cycle, ultimately increasing noise in the manifold generation process and corrupting the motion modeling. These methods were also primarily designed either for gating purposes or not designed for MRI specifically, which exhibits several challenges in terms of spatial and temporal resolution. Furthermore, some proposed ML-based methods have been validated either on synthetic data or on low temporal resolution data [5,71–73]. In these cases, the discrimination between slices is simplified as it uses images of high quality.

In feature extraction methods, the derivation of a reliable respiratory signal from the acquired images is used to optimize the reordering process. Some prior work has proposed to monitor the body area to represent the breathing signal as it typically correlates with the breathing motion [6, 76, 77]. In [6] and [76], 2D dynamic axial images were used for the reordering process. However, axial planes are rarely used since abdominal motion is better appreciated in the sagittal and coronal planes. In [77], it was demonstrated that sagittal slices yield more accurate 3D volume reconstructions. Nevertheless, changes of body area are prone to be affected by space-dependent phase shifts. The approaches presented in [75] and [78] are based on calculating the mutual information between slices. Unfortunately, slice reordering methods based only on imaging data may not guarantee adequate temporal behavior. Image-based internal surrogates using dimensionality reduction has been proposed in previous work [79]. However, the main limitation of this approach is the lack of a direct relationship between the surrogate signal and organ motion: the low-dimensional representation of the images does not always proportionally change with respiratory motion. Van de Lindt, Sonke, Nowee, Jansen, van Pelt, van der Heide and Fast [68] proposed to bin 2D slices according to their craniocaudal motion to construct 4D volumes. However, the validation is performed against a navigator signal, thus limiting the temporal resolution of their 2D slice series acquisition and makes the binning process easier. Moreover, the navigator, a craniocaudal 1D signal cutting the right hemi-diaphragm at a given position, is insensitive to liver deformation that occur far from its position.

The closest work to ours was presented by Tong, Udupa, Ciesielski, Wu, McDonough, Mong and Campbell [199]. The authors propose to construct a weighted graph and to reconstruct volumes following the shortest paths. However, such method is limited to sequences where lungs are entirely visible. Moreover, it assumes that the respiratory signals are regular and a manual selection of the end-exhale positions for each slice is required. As a result, the manual inspection of about 10 000 images per patient would be required for our application. We propose an automatic weighted graph-based method, using both image and motion information, that is able to handle irregular respiratory motion and that can even be applied to images in which the lungs are partially visible. The method is tested on the liver in comparison to a slightly altered version of [199] and a state of the art method using ML [73]. Finally, we propose new temporal metrics to assess the quality of slice reordering regarding the temporal coherence of the reconstructed volumes.

## 5.3 Material and Methods

The proposed slice reordering method is based on inter-slice similarity measures which take into account both pixel and motion information (see Figure 5.1). These similarity measures are used to construct a weighted graph $\mathcal{G}$ where each vertex $\mathcal{V}_{s,t}$ represents $(s,t)$, the slice $s$ at time $t$. The edges of the graphs connect only neighboring slices: the edge $\mathcal{E}_{s,t,t'}$ links vertex $\mathcal{V}_{s,t}$ to vertex $\mathcal{V}_{s+1,t'}$ where $s \in \{1,\ldots,N_s\}$ and $t,t' \in \{1,\ldots,N_t\}$; $N_s$ is the number of slices and $N_t$ the number of time points for each slice. Each vertex $\mathcal{V}_{s,t}$ is associated with the weight $w_{s,t,t'}$ computed from the inter-slice pixel and motion similarity measures. The volumes are reconstructed by finding the shortest paths in $\mathcal{G}$.

Image similarity ensures spatial consistency between neighboring slices and is reflected in two weighting terms: an inter slice pixel wise image similarity measure and right hemi-diaphragm height consistency. Motion information is derived from the automatic computation of a pseudo navigator $n_s(t)$ for each slice $s$. This differs from previous approaches [199] where only manually labelled end-exhale images are used.

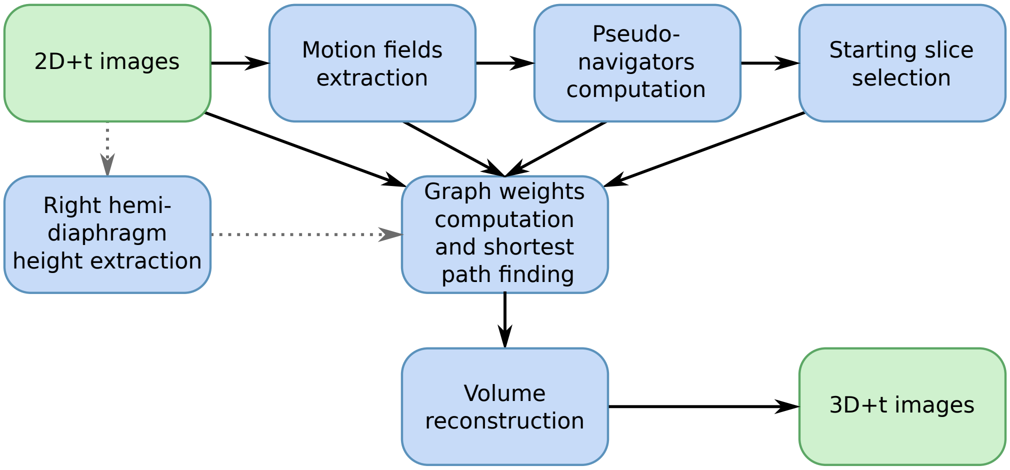

2D+t images
Motion fields extraction
Pseudo-navigators computation
Starting slice selection
Right hemi-diaphragm height extraction
Graph weights computation and shortest path finding
Volume reconstruction
3D+t images

Figure 5.1 Overall scheme of the proposed method. The core of the method is based on both image and motion metrics to construct a weighted graph. Motion fields are extracted from the 2D dynamic slice series and used to compute a pseudo-navigator for each slice which describes the respiratory motion. The weight quantifying the degree of coherence of the right hemi-diaphragm position across slices is set as an option as this weight is specific to our dataset. The volume reconstruction is performed from a reference slice, selected based on characteristics of its pseudo navigator, toward the first and last sagittal slices following the shortest path in the graph.

### 5.3.1 Data acquisition

Free breathing high-resolution sagittal slices were acquired on twelve volunteers, who provided their written consent. The acquisitions were carried out on a Siemens Skyra 3T scanner using a 2D T2-weighted true FISP sequence with a pixel spacing of $1.7 \times 1.7$ mm$^2$ and a slice thickness of 3 mm. To cover the whole liver, between 66 and 84 slices were acquired, depending on the liver size. Each slice position was imaged 150 times, for a total of 20 seconds, which covers approximately 4 to 6 respiratory cycles, without any gating method.

### 5.3.2 Automatic pseudo-navigator extraction

In MRI acquisitions, the term navigator refers to a one dimensional signal located at the summit of the right hemi-diaphragm dome. This signal provides the height of the diaphragm at the dome summit and is generally used to perform respiratory gated MRI acquisitions. Here, we use the term pseudo-navigator to designate a computed signal that gives a relative diaphragm height for each slice position. The automatic extraction of this pseudo navigator signal is performed for each slice $s$ as follows (see Figure 5.2):

First, the displacement field between $(s, t)$ and $(s, t+1)$ is computed for all $t \in \{1, \dots, N_t\}$, and its vertical component median value $\bar{v}_s^{\text{raw}}(t)$ is extracted with the convention that the positive direction is upward. Dense displacement fields between two temporal slices were calculated using the NiftyReg software [201]. Specifically, cubic B-Splines transformation were generated to deform a source image in order to optimize an objective function based on the Normalized Mutual Information and a penalty term based on the bending-energy [30].

Then, a low pass filtering with a cutoff frequency of 0.5 Hz is applied on $\bar{v}_s^{\text{raw}}(t)$ to remove the noise and keep only the respiratory signal, giving the signal $\bar{v}_s(t)$. Zero crossing points of $\bar{v}_s(t)$ with negative derivative correspond to end-exhale positions.

Finally, the first end-exhale slice $(s, t_{E_0}^s)$ where $t_{E_0}^s$ is the time point corresponding to the first end-exhale of slice s, is selected to serve as reference and the navigator signal $n_s(t)$ is computed as follows:

$$n_s(t) = \int_{t_{E_0}^s}^{t} \bar{v}_s(x)dx \tag{5.1}$$

### 5.3.3 Right hemi-diaphragm height extraction

The diaphragm position is found before reordering for all images $I_{s,t}$ with $s \in \{s_L, ..., s_R\}$ with $s_L$ and $s_R$ the most left, respectively right, sagittal slice index for which the detection is performed and correspond to the limits of the right hemi-diaphragm and $t$ the time index

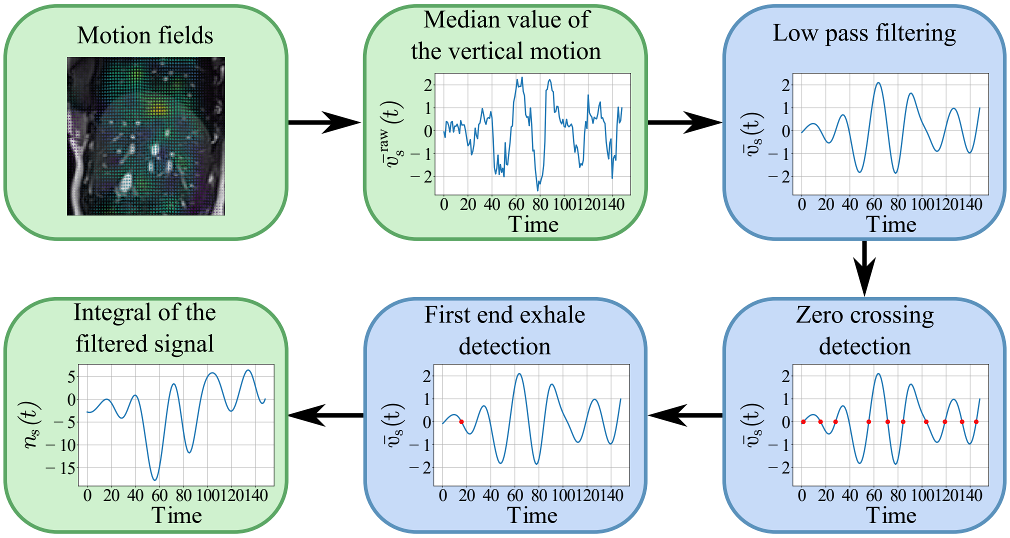


Figure 5.2 Navigator extraction process.

before reordering. The diaphragm position is found by gradient based edge detection along the superior-inferior (SI) axis. First, $I_{s,t}$ is cropped and the maximal intensity value saturated at the average liver pixel intensity to avoid noisy signal from the bowel area and the liver vessels. The resulting image $I_{s,t}^{\text{c,sat}}$ is then multiplied pixel-wise by a parabolic shading image in order to attenuate bright signal coming from the lung airways: $I_{s,t}^{\text{c,sat,shaded}} = I_{s,t}^{\text{c,sat}} \cdot I^{\text{para}}$. A Gaussian filtering followed by binary threshold around lung pixel mean intensity (90 in this case) is applied on $I_{s,t}^{\text{c,sat,shaded}}$ to have a rough segmentation of the visible part of the right lung: $L_{s,t}$. Morphological 7x7 filters are used to include all airways in the rough segmentation: $L_{s,t}^{\text{morph}}$. A Gaussian gradient magnitude filter is applied on both $I_{s,t}^{\text{c,sat,shaded}}$ and $L_{s,t}^{\text{morph}}$ before multiplying them pixel wise to obtain the final gradient image. The maximum gradient intensity is computed column wise and its position is stored as the diaphragm height of this column: $\Delta_{s,c}^{\text{raw}}(t)$ (see Figure 5.3).

The final diaphragm height $\Delta_{s,c}(t)$ before reordering is obtained after a temporal and spatial smoothing of $\Delta_{s,c}^{\text{raw}}(t)$ to correct aberrant values. The temporal smoothing uses the fact that before reordering the difference in diaphragm height for column $c$ between $(s,t)$ and $(s,t+1)$ is small while the spatial smoothing in sagittal slices take advantage of the fact that the liver is a smooth organ. During this process a confidence value $\mathcal{C}_{s,c,t}$ is associated to each computed height.

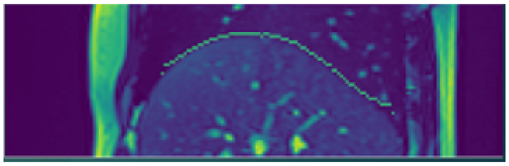


Figure 5.3 Detected diaphragm points in a cropped sagittal slice.

### 5.3.4 Edge weight computation

The total weight $w_{s,t,t'}$, associated to each edge $\mathcal{E}_{s,t,t'}$ is composed of several weighting terms. The first component, $w_{i_{s,t,t'}}$, reflects the image similarity between slices $(s,t)$ and $(s+1,t')$. Before the similarity is computed, in order to remove flashing artifacts caused by the blood vessels of the liver, image intensities are saturated at an intensity value $I_{\text{sat}}$, set to approximately the mean intensity of the liver. Further, the images are cropped so that the lower abdomen is left out, as the changes in this region have low correlation with the respiratory induced motion. By visual inspection we found that in our dataset the liver always appears centered in the superior part of the image. Therefore, we extracted this ROI using the knowledge on the spatial distribution of the organ within the images. The weighting term is computed as:

$$w_{i_{s,t,t'}} = \frac{1}{N_p} \sum_{(x,y)} \left( \frac{I_{s,t}(x,y) - I_{s+1,t'}(x,y)}{0.5 \times I_{\text{sat}}} \right)^2 \tag{5.2}$$

where $N_p$ is the number of pixels in the images after cropping and $I_{s,t}(x,y)$ is the intensity of pixel $(x,y)$ of slice $s$ at time $t$ after saturation.

The second component, $w_{p_{s,t,t'}}$, is a measure of the difference in phase on the respiratory cycle between slices $(s,t)$ and $(s+1,t')$, defined as $d = \left| \frac{k_{s,t}}{l_{s,t}} - \frac{k_{s+1,t'}}{l_{s+1,t'}} \right|$. Here $k_{s,t}$ denotes the number of time points between $t$ and the last end-exhale time point preceding $t$ in slice $s$ and $l_{s,t}$, the respiratory cycle length, denotes the number of time points between the last end-exhale preceding $t$ and the first end-exhale after $t$. The weight is given by:

$$w_{p_{s,t,t'}} = \frac{1 - \exp^{-d}}{1 - \exp^{-1}} \tag{5.3}$$

The third component, $w_{a_{s,t,t'}}$ is a measure of the amplitude difference between the respiratory cycles pseudo navigators in which $t$ and $t'$ belong to:

$$w_{a_{s,t,t'}} = 1 - \exp^{-(A_{s,t} - A_{s+1,t'})^2} \tag{5.4}$$

where $A_{s,t}$ is the amplitude of the pseudo navigator signal of slice $s$ on the respiratory cycle that $t$ belongs to.

The fourth component $w_{n_{s,t,t'}}$ is a measure of the difference in the relative height of the pseudo navigators:

$$w_{n_{s,t,t'}} = 1 - \exp^{-(\frac{n_s(t)-n_s(t_I)}{A_{s,t}} - \frac{n_{s+1}(t)-n_{s+1}(t'_I)}{A_{s+1,t'}})^2} \tag{5.5}$$

where $t_I$, respectively $t'_I$, is the inhale time-point of the respiratory cycle $t$, respectively $t'$, belongs to.

The last component, $w_{r_{s,t,t'}}$, determines the slice motion difference between $(s+1,t)$ and the reference slice $(s_r, t_i)$ chosen to start the reordering process:

$$w_{r_{s,t,t'}} = D_\kappa\left(\frac{\bar{v}_{s_r}(t_i)}{l_{s+1,t'}}, \frac{\bar{v}_{s+1}(t')}{l_{s_r,t_i}}, \operatorname{sgn}\left(\dot{\bar{v}}_{s_r}(t_i)\right), \operatorname{sgn}\left(\dot{\bar{v}}_{s+1}(t')\right)\right) \tag{5.6}$$

where $D_\kappa$ is a weighted distance: $D_\kappa(a,b,c,d) = \sqrt{(a-b)^2 + \kappa(c-d)^2}$, $\operatorname{sgn}(x)$ is 1 if $x$ is positive and -1 if $x$ is negative, and $\dot{}$ denotes the first derivative. This weighting term avoids cumulative shifts between the vertical component median value of the motion field across the slices.

Finally, the weight $w_{s,t,t'}$ of edge $\mathcal{E}_{s,t,t'}$ is computed as follow:

$$w_{s,t,t'} = \alpha w_{i_{s,t,t'}} + \beta w_{p_{s,t,t'}} + \gamma w_{a_{s,t,t'}} + \delta w_{n_{s,t,t'}} + \epsilon w_{r_{s,t,t'}} \tag{5.7}$$

where $\alpha$, $\beta$, $\gamma$, $\delta$, and $\epsilon$ are scalar factors balancing the importance of each term.

In this application using liver images we included an additional weight: $w_{d_{s,t,t'}}$.This weighting term ensures the coherence in the right hemi-diaphragm position across adjacent slices located between $s_R$ and $s_L$. The right hemi-diaphragm height $\Delta_{s,c}(t)$ is automatically extracted for each column $c$ of $s$, see Section 5.3.3. The weight $w_{d_{s,t,t'}}$ is expressed as follows:

$$w_{d_{s,t,t'}} = |(\bar{\Delta}^E_{s,c} - \bar{\Delta}^E_{s+1,c}(t')) - (\Delta_{s,c}(t) - \Delta_{s+1,c}(t'))| \tag{5.8}$$

where $\bar{\Delta}^E_{s,c}$ is the median value of the diaphragm height of all end-exhale time points for column $c$ of slice $s$. This weighting term can be omitted or changed depending on the application.

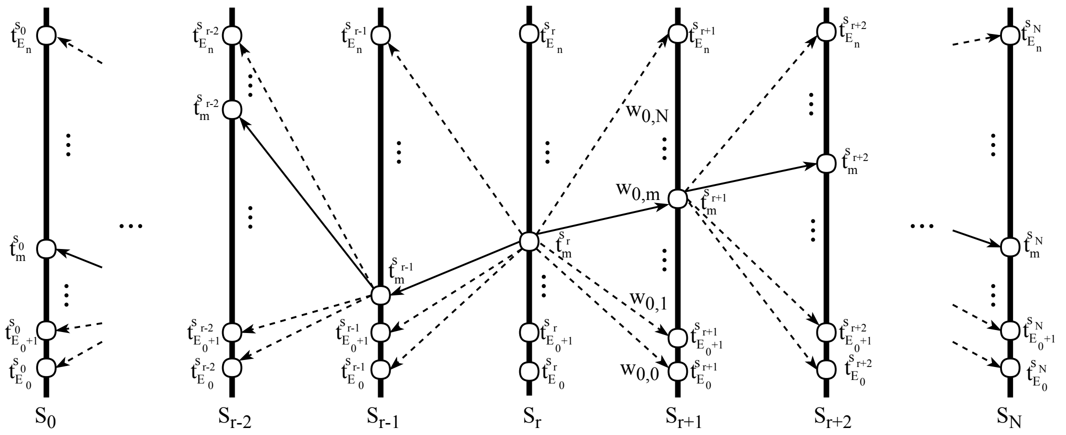


Figure 5.4 Slice matching algorithm. Each vertical line represents an anatomical position $S$ which is acquired $t_N$ times before acquiring the next one. The slice matching process starts from a reference slice $S_r$. The reconstruction is done in 2 directions: backward and forward, as the arrows show. One vertex $(s_r, t_i)$ is connected to all the vertices in the next slice (dashed lines). Each edge is given a weight $w_s, t$ to measure the matching unlikeness between the 2 compared slices. The vertex with less value is considered the matching time $t_{m_{S_{r+1}}}$ (solid line). This is repeated up to construct all the volumes.

### 5.3.5 Volume reconstruction

Volumes are reconstructed by determining the optimal paths on the graph (the ones associated with the smallest weight) starting at each time step between the first end-exhale, $t_{E_0}$, and the last end-exhale, $t_{E_n}$, of a reference slice $s_r$ (see Figure 5.4). The volume at $t_i$ is reconstructed from vertex $(s_r, t_i)$, following the optimal paths on the graph in two directions: toward the first slice and toward the last slice. This process is repeated for all time steps to obtain a sequence of constructed volumes. The reference slice $s_r$ is selected based on motion criteria. Selection of the starting slice $s_r$ is made when a slice satisfies the following requirements: $s_r$ cuts the right hemi-diaphragm, breathing motion should not halt over the period covered by $s_r$, $s_r$ spans over at least one complete respiratory cycle, and the amplitude of $n_{s_r}$ is the median value of the amplitude of $n_s$ for all slices under the right hemi-diaphragm.

## 5.4 Validation metrics

### 5.4.1 Inter-slice diaphragm consistency

An inter-slice diaphragm metric is used as a spatial metric by considering the set of slice positions $s \in \{s_L, ..., s_R\}$ for which the diaphragm position has been found (see Section

5.3.4). A coronal slice position $c$, where the right lobe of the liver is well visible, is manually selected for each volunteer for validation purpose only. Subsequently, the mean of the sum of the diaphragm position variance inside a sliding window, $\mu_c(t^a)$, is computed: $\mu_c(t^a) = \frac{1}{3(M-1)}\sum_{i=s_L}^{s_R-1}\sum_{j=-1}^{1}(\Delta^a_{i+j,c}(t^a) - \hat{\Delta}^a_{i,c}(t^a))$, where $M = s_R - s_L + 1$ is the number of slice positions with a diaphragm signal, $\Delta^a_{i+j,c}(t^a)$ is the diaphragm height after reordering of the column $c$ of sagittal slice $i + j$ in the reconstructed volume at time $t^a$ and $\hat{\Delta}^a_{i,c}(t^a)$ is the mean value along the coronal direction of the diaphragm position inside the sliding window centered at the column $c$ of slice $i$ at time $t^a$. $\mu_c(t^a)$ is calculated for all time points and the sum is the diaphragm discrepancy metric:

$$d_{DH}(t^a) = \sum_{\text{volunteers}} \mu_c(t^a) \tag{5.9}$$

### 5.4.2 Temporal metrics

Using the original time positions of the slices chosen in the reconstruction, a new pseudo-navigator signal for each sagittal slice series was constructed. The navigator signal after reconstruction was obtained from the values computed before reordering replaced at their new time point. For instance, if slice $(s,t)$ has been stacked to the volume at $t^a$, the pseudo navigator value $n^b_s(t)$ computed for $(s,t)$ before reordering will be used for slice $(s,t^a)$ in the reordered volume: $n^a_s(t^a) = n^b_s(t)$. Finally, a new set of signals, $\nu^a_s(t^a)$ is obtained by smoothing the pseudo-navigator signals using an amplitude conserving filter.

The distance $d_{n_s}(t^a)$ is defined as the absolute difference between the original reconstructed pseudo-navigator and the corresponding smoothed signal:

$$d_{n_s}(t^a) = |n^a_s(t^a) - \nu^a_s(t^a)| \tag{5.10}$$

The distance $d_{n_s}(t^a)$ allows for the detection of misaligned slices as they correspond to non-smooth regions of the post construction navigator curve.

An inter-slice comparison is performed by calculating the difference of the filtered signal $\nu^a_s(t^a)$ at each slice position with the signal of the reference slice $s_r$, $\nu^a_{s_r}(t^a)$.

$$d^{\intercal}_{n_s}(t^a) = |(\nu^a_{s_r}(t^a) - \bar{\nu}^a_{s_r}) - (\nu^a_s(t^a) - \bar{\nu}^a_s)| \tag{5.11}$$

where $\bar{\nu}^a_s$ is the mean of the filtered signal after reordering for slice $s$. The distance $d^{\intercal}_{n_s}(t^a)$ for each slice $(s,t^a)$ enables stacking discrepancy detection using the reference slice navigator signal.

Furthermore, the computation of the maximum amplitude $A^s_{\max}$ for each navigator signal $\nu_s$ enables the detection of slices that show less motion than the reference slice. $A^s_{\max}$ is calculated as the dynamic range of the linearly detrended pseudo navigator.

Finally, estimates of the number of breathing cycles present in each signal are computed using the reconstructed navigators. Assuming that the navigators are centered around zero and carry a dominant, low frequency, sinusoidal signal, they are low pass filtered and linearly detrended, and such a function is found using least squares curve fitting, solving the following minimization problem: $\arg\min_{A_s,\omega_s,\theta_s} \sum_{t^a=1}^{N_t}(n^a_s(t^a) - A_s \sin(\omega_s t^a + \theta_s))^2$, where the frequency, $\omega_s$, is used to calculate the number of oscillations in the signals, corresponding to the number of detected breathing cycles, $c_s$, at each slice position. It is postulated that the total number of breathing cycles should be consistent across all slices after reconstruction, and in particular, that they should be consistent with the number of cycles detected in the reference slice so that $c_{s_r} \approx c_s,\ \forall s \in \{1, ..., N_s\}$.

## 5.5 Results

### 5.5.1 Validation of the automatic end-exhale extraction

The automatic end-exhale detection algorithm was validated against a manual labelling (see Table 5.1). There are 3 groups of slices which show similar motion patterns and appearance: (1) liver slices which are located spatially before the cardiac cavity, (2) slices where the cardiac cavity appears and only a small portion of the liver, (3) slices where only the cardiac cavity appears without the liver. Because the liver appearance can be quite different between these three scenarios when sweeping the liver volume, we decided to validate the end-exhale detection algorithm using one exemplar from each representative group. This allows to evaluate the accuracy from different spatial locations within the liver.

We would like to clarify the breathing during the exhalation phase can induce motion dynamics in the liver. At the beginning of the exhale phase, the liver progressively goes up towards the diaphragm until it reaches a maximum height. Then, there is a short resting period in which it remains in the same height before start the inhalation phase. The manual annotation of the first end-exhale was performed by five operators with background in medical imaging or radiology and several years of experience in the field. They were given precise instructions to identify the last temporal slice where the liver achieves the highest height (in the first respiratory cycle). In other words, the temporal slice before the liver starts to descend.

In general, the manual annotation of the end-exhale phase is a time-consuming, tedious and prone to errors task. In data acquired with high temporal resolution it is unfeasible the labeling of thousands of images by one operator. Moreover, there are some factors which increases the difficulty of manual end-exhale extraction: prolonged apnea, deep and shallow breaths and in general high inter-fraction variability. Several observations can be made from the results presented in Table 5.1. First, we can see that overall there is a good agreement between the manual and the automatic labelling; moreover, the standard deviation (stdev) is generally larger for the most difficult cases, on the slice (S2), where the heart is visible. Two cases are of particular interest: slice S2 of volunteer 11 and slice S3 of volunteer 1. In both there is a strong disagreement between the different operators. This is due to the fact that one of the operator missed the first end-exhale and another one the two first end-exhales, showing the difficulty of manual labelling. Secondly, the likelihood to miss end-exhale positions will depend on the inter-fraction variability showed by the subject. In our dataset we observed cases where it was extremely difficult to segment the respiratory cycles because the liver remains for more than 10 temporary slices in the same position and/or does not descend up to the same height. Finally, the spatial location of the slice also influences this process. In anatomical positions like the one that crosses the cardiac cavity, the level of uncertainty in the detection increases due to the influence of the cardiac motion over a small liver portion. Thus the detection is challenging even for expert operators. This demonstrates the relevance of the proposed automatic end-exhale detection algorithm.

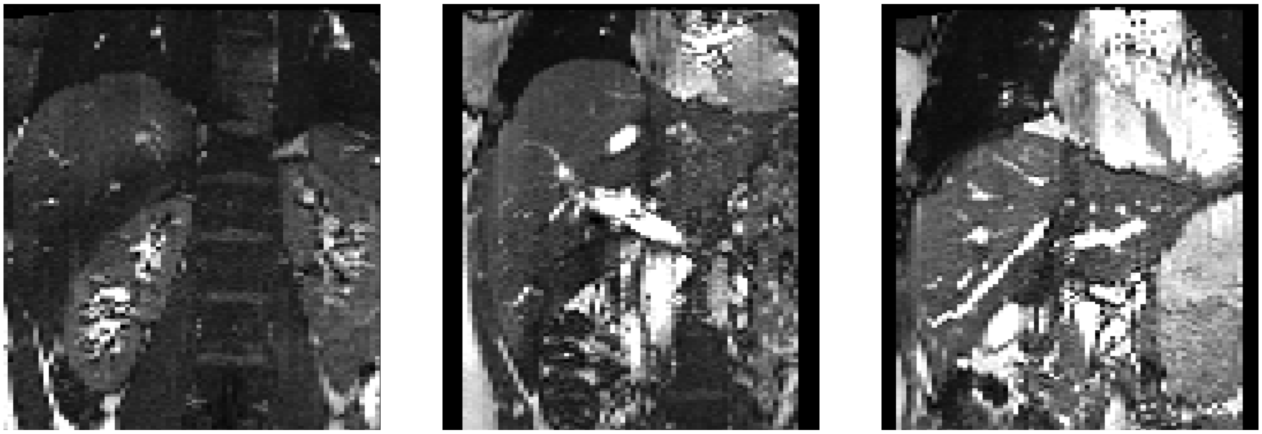
The proposed slice reordering method for 4D image construction was compared against two state of the art manifold alignment (MA) [73] and feature extraction [199] based techniques respectively on a dataset of 12 volunteers. All validated using a set of spatial and temporal metrics. The MA implementation uses Wave Kernel Signature (WKS) as graph descriptor. Fully connected graphs were used with $\sigma_G = 1.5$ as suggested for the authors when using image intensities as high-dimensional data. The parameters related with the WKS descriptor were $\sigma_{WKS} = 0.8$ and $\mu = 0.9$. We performed the following three modifications in [199] to be able to use it in our data. First, we used our automatic method to identify the exhale positions within each slice. Since we had an average of 10,000 images per volunteer, it was unfeasible to perform a manual labeling. Secondly, the lungs were not entirely visible on our dataset; in turn, we cropped the bottom part of our images to increase the proportion of the lung in the images. This way, we calculate the similarity between images containing the upper part of the liver, the diaphragm and a portion of the lungs. Lastly, for consistency we selected the same starting slice $s_r$ for all three methods. In their article Tong, Udupa, Ciesielski, Wu, McDonough, Mong and Campbell [199] always select the first slice as their starting slice, but this choice greatly impairs the performance of the method on our dataset.

Table 5.1 Automatically extracted first end-exhale time point indices in three different slice positions for all subjects compared to manually selected time points (median ± stdev) of 5 operators. The three slice positions correspond to an area covering the liver and right hemi-diaphragm (S1), the heart (S2) and the left hemi-diaphragm (S3). Finding end-exhale time points on S1 and S3 is considered relatively easy, while on S2 it is challenging because of the heart motion.

| | S1 | | S2 | | S3 | |
|---|---|---|---|---|---|---|
| | Auto. | Manual | Auto. | Manual | Auto. | Manual |
| V1 | 19 | 20 ± 2.9 | 14 | 16 ± 4.6 | **7** | **9** ± **25.0** |
| V2 | 23 | 20 ± 2.6 | 16 | 19 ± 4.7 | 15 | 19 ± 4.2 |
| V3 | 27 | 27 ± 3.2 | 19 | 20 ± 1.9 | 12 | 13 ± 1.5 |
| V4 | 23 | 24 ± 2.5 | 10 | 16 ± 4.8 | 23 | 25 ± 1.5 |
| V5 | 25 | 26 ± 1.4 | 17 | 20 ± 3.2 | 28 | 26 ± 1.4 |
| V6 | 33 | 32 ± 2.4 | 38 | 32 ± 6.5 | 21 | 22 ± 3.8 |
| V7 | 14 | 18 ± 2.4 | 19 | 21 ± 4.3 | 9 | 11 ± 2.3 |
| V8 | **21** | **36** ± **5.7** | 23 | 14 ± 8.4 | 17 | 20 ± 1.0 |
| V9 | 35 | 37 ± 4.4 | 32 | 27 ± 6.6 | 25 | 24 ± 4.1 |
| V10 | 16 | 17 ± 1.7 | 10 | 9 ± 0.4 | 10 | 11 ± 6.4 |
| V11 | 58 | 58 ± 6.1 | 15 | 17 ± 13.6 | 48 | 50 ± 5.7 |
| V12 | 6 | 9 ± 6.4 | 23 | 26 ± 1.7 | 19 | 21 ± 1.4 |

### 5.5.2 Evaluation of the proposed 4D construction method

For our method two sets of parameters were used. One with all set to 1.0: $\alpha = 1.0$, $\beta = 1.0$, $\gamma = 1.0$, $\delta = 1.0$, and $\epsilon = 1.0$, and the other disregarding image similarity by setting $\alpha = 0.0$



(a) Volunteer 3, slice 116 at time point 2.

(b) Volunteer 5, slice 91 at time point 93.

(c) Volunteer 9, slice 78 at time point 76.

Figure 5.5 Coronal view of the reconstructed volume of 3 volunteers.

while keeping all other set to 1.0. Further, it was found that the method in general was insensitive to the choice of these weighting scale factors. These weighting scale factors were defined systematically using heuristic trial and error while evaluating its influence on metric results.

Figure 5.5 shows a qualitative result of the 4D construction method on 3 volunteers. The spatial consistency is very good for the part of the liver under the right hemi-diaphragm even if for these result the optional weight $w_d$ was not used. In the part under the heart some slices are misaligned. In fact, it is difficult to ensure that the 2D acquisition has captured all the possible combination of the heart and respiratory motions. The slice stacking may thus be suboptimal because of the missing information.

Figure 5.6 shows the average variance of the diaphragm height (see 5.4.1) in all volunteers as a function of time for each reordering method. The lower values, which were obtained with the proposed method, indicate a greater spatial consistency.

The result of the validation using the two navigator based temporal metrics (see Section 5.4.2) across all volunteers is presented in Figures 5.7 and 5.8. It is shown for both sets of parameters for our method, one using combined image similarity and motion information and one using only motion derived weights, as well as the two other methods. Overall, it can be seen that our proposed method outperforms the other two, showing smoother reconstructed navigator signals and lower inter-slice dissimilarities. The proposed method was a significant improvement on the compared methods ($p$-value=0.002 in both cases) for the intra-slice reconstructed navigator smoothness. However, statistical analysis showed that the proposed method did not provide a significant improvement on the inter-slice reconstructed navigator

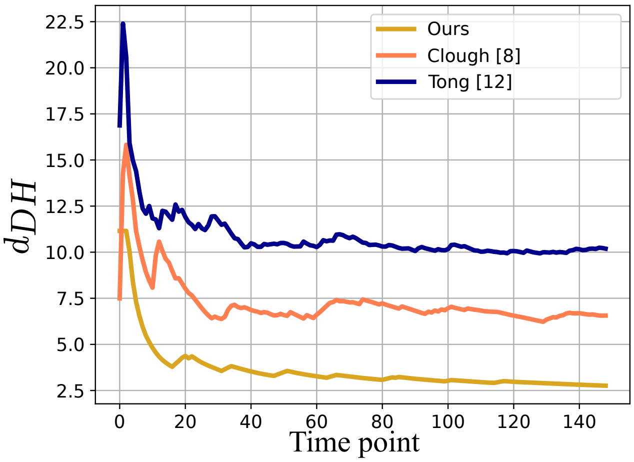


Figure 5.6 Diaphragm height discrepancy

dissimilarity ($p$-value=0.2). This may be because the test was done with insufficient number of cases. Statistical analyzes were performed using functions implemented in the Python Scipy library. Wilcoxon signed-rank test was used since the data is not normally distributed.

The same metrics are shown for volunteer 5 in Figures 5.9 and 5.10 for all slice positions and time points where each row corresponds to the validation of one constructed navigator signal. We can see that in this case our method produces smoother reconstructed navigators over the whole 4D sequence, Figure 5.9, as well as having lower inter-slice differences across the liver. Further, it can be seen in Figure 5.11 that the proposed method has reconstructed navigator signals of appropriate amplitudes compared to the ones before construction, showing more motion across the right hemi-diaphragm and a near consistent number of estimated cycles across the whole imaged region.

Figure 5.12 shows an analysis of each single weight contribution. The inter-slice validation result is similar for all thus improvements concerning the intra-slice metric is more interesting to analyze. For the intra-slice reconstructed navigator smoothness, the best results were yielded with $w_a$ and $w_n$. These weights have shown to be the most influential in the volume construction. Slice matching that belong to respiratory cycles with similar amplitude provides robustness to the volume construction since it copes with the irregular breathing.

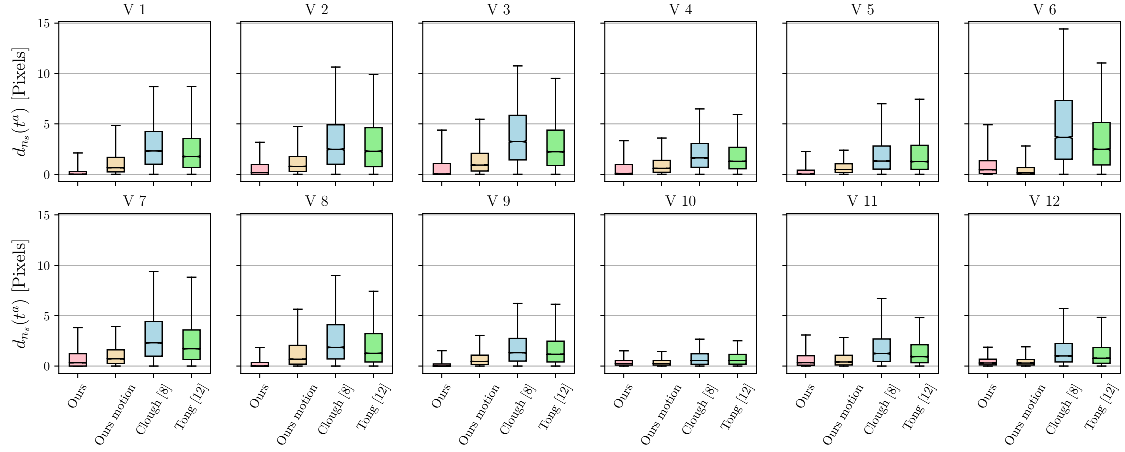


Figure 5.7 Intra-slice reconstructed navigator smoothness, a lower value is better. Box plot shows the minimum value, the first quartile, the median, the third quartile and the 95$^{th}$ percentile of $d_{n_s}(t^a)$ for all time points and all slice positions.

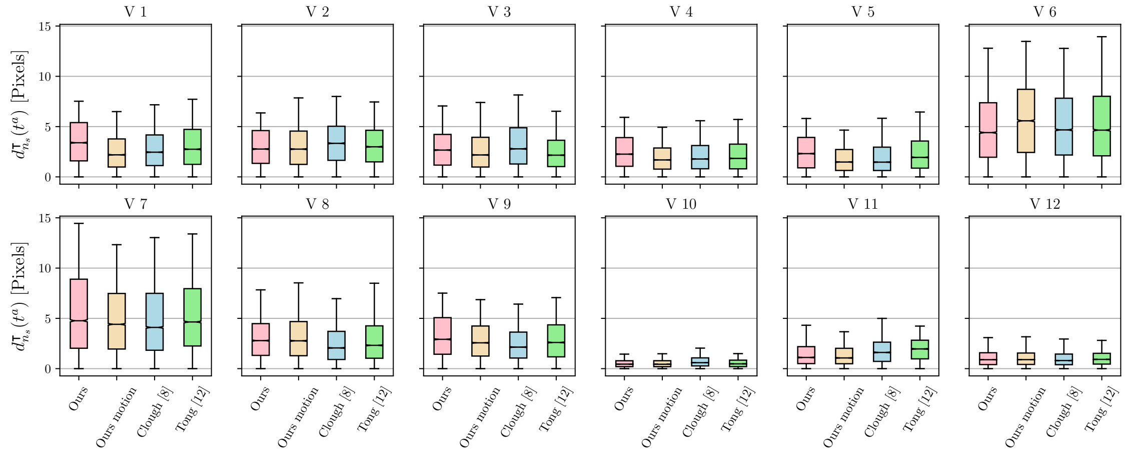


Figure 5.8 Inter-slice reconstructed navigator dissimilarity. Box plot shows the minimum value, the first quartile, the median, the third quartile and the 95$^{th}$ percentile of $d^{\intercal}_{n_s}(t^a)$ for all time points and all slice positions.

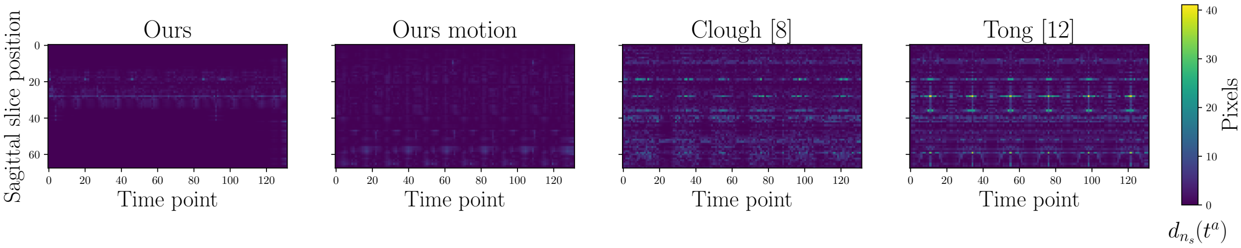


Figure 5.9 Intra-slice temporal metric, based on Eq. 5.10, for all time points and all slices for volunteer 5, showing near zero values for our proposed method.

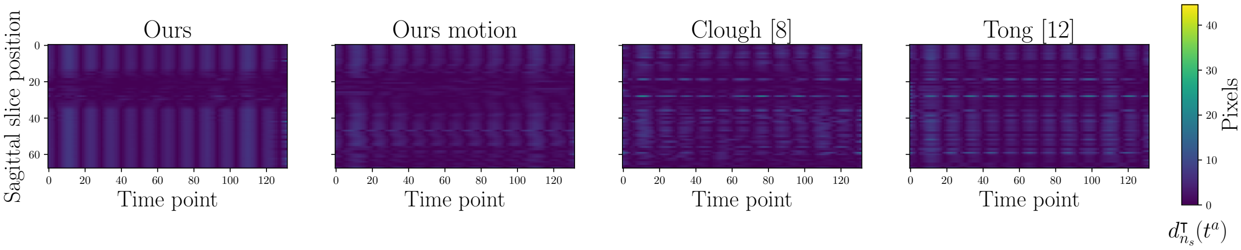


Figure 5.10 Inter-slice temporal metric, Equation 5.11, for all time points and all slices for volunteer 5. Slice positions 10 to 40 correspond roughly to a section covering the liver and the right hemi-diaphragm.

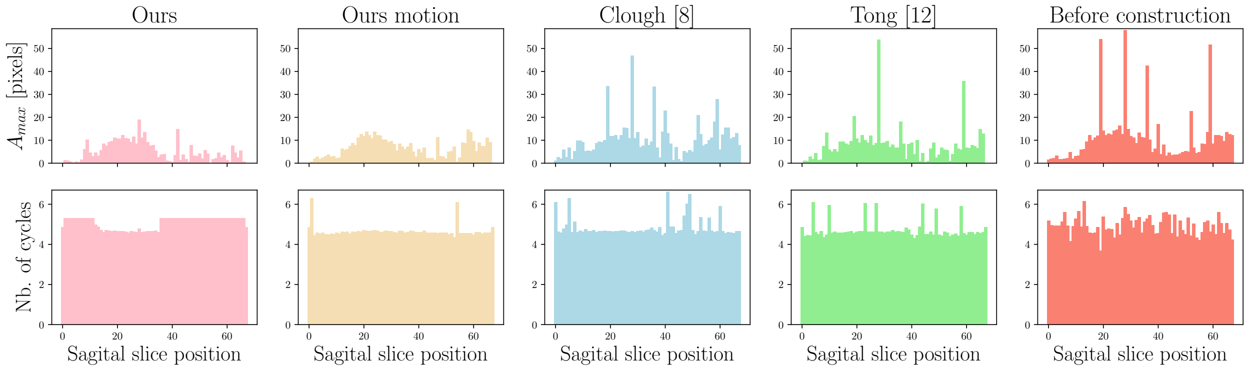
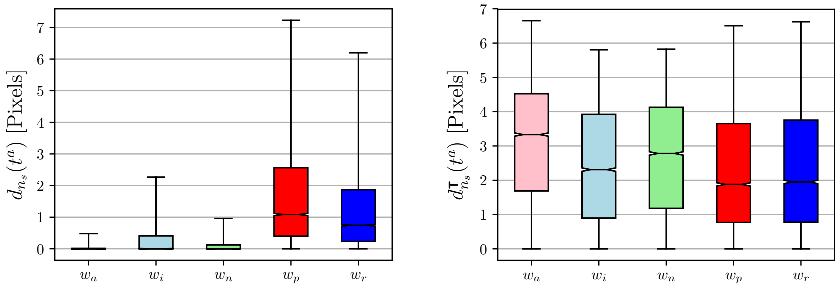


Figure 5.11 Navigator amplitude and estimated number of cycles before and after construction for volunteer 5.



(a)

(b)

Figure 5.12 Contribution of each single weights for the construction (a) Intra-slice reconstructed navigator smoothness (a lower value is better). (b) Inter-slice reconstructed navigator dissimilarity.

## 5.6 Discussion and conclusion

The results presented here demonstrate that the proposed method outperforms state of the art methods for slice reordering both in terms of spatial and temporal quality. Compared to other methods that assume a regular respiratory pattern, our method is able to cope with irregular breathing and with small breath-hold of the volunteers. The automatic end-exhale time detection and the automatic pseudo navigator extraction allow the method to work on high spatial and temporal resolution data that capture several respiratory cycles, enabling inter-cycle variability studies. The flexibility in starting slice selection enables the reconstruction of different 4D image sequences and increase the variability of respiratory motion patterns that an operator can choose to capture. This is especially interesting for motion model construction.

The proposed method does not use any information from the previous sorting during the current sorting, in order to avoid propagating potential inconsistencies. Nevertheless, it may occur that some slice stacked during the construction of one volume at particular point in time may be used for others volumes as long as it gets the lowest weight value compared with the all the remaining slices. In the practice, this means that the 4D image construction is a numerical approximation of the real organ motion.

In this study, the weighting parameters were chosen empirically. Surprisingly, the image similarity weight did not seem to have any positive effect on the 4D construction in our method, suppressing motion by favoring the slice from the same time point throughout the volume sequence. As part of our future work, a grid search will be implemented to optimize the weighting parameters; we expect an increase of the reconstruction accuracy. From a visual inspection of the reconstructed volumes with the 3 methods, it can be concluded that slice stacking at inhalation positions is more difficult because the liver does not always descend to the same position.

To accurately reconstruct the area under the heart, it is important to ensure that most of the possible combinations of heart and respiratory motions are acquired for all the slices. Alternatively, the heart motion can be suppressed using cardiac gating; however, this alternative will reduce the temporal resolution and the respiratory motion may not be thoroughly acquired. A final possibility would rely on an algorithm which can remove cardiac motion using a golden-angle radial acquisition or post-processing compensation tools.

One of the main limitations of existing methods, including the one presented in this paper, is the absence of a global temporal consistency measure during optimization. The main challenge with this optimization is the large increase in computational time required. However, we believe that finding a way to add global temporal coherence in a slice reordering method would greatly improve the quality of the results and should be investigated in future work.

**Funding:** This work was partly funded by an NSERC collaborative research and development project (CRDPJ-517413-17).

**Conflict of interest:** The authors declare that they have no conflict of interests.

**Informed consent:** Informed consent was obtained from all individual participants included in the study.

# CHAPTER 6 ARTICLE 2: PREDICTIVE ONLINE 3D TARGET TRACKING WITH POPULATION-BASED GENERATIVE NETWORKS FOR IMAGE-GUIDED RADIOTHERAPY

Contribution of the first author in preparation and writing this paper is evaluated as 90%. This article has been published by the International Journal of Computer Assisted Radiology and Surgery on June 2021. Additionally, this work won the *Best Paper Award* in the $12^{th}$ International Conference on Information Processing in Computer-Assisted Interventions.

**Remarks:** This paper presents the first population-based deep model able to relate temporal 2D slices with a future 3D deformation. It introduces a methodology to relate input images and dense deformations by minimizing their distances within a low-dimensional space. It also demonstrates that the created latent space in meaningful since it contains respiratory phase information. The results show the capability of the network to predict future anatomical changes while bringing important advantages over existing techniques.

**Predictive online 3D target tracking with population-based generative networks for image-guided radiotherapy**

Liset Vázquez Romaguera[1], T. Mezheritsky[1], Rihab Mansour[2], William Tanguay[3], Samuel Kadoury[1,2]

[1] Polytechnique Montreal, Canada, [2] Centre Hospitalier de l'Université de Montréal Research Center, [3] Département de Radiologie, Radio-Oncologie et Médecine Nucléaire. Faculté de médecine, Université de Montréal.

## 6.1 Abstract

*Purpose* Respiratory motion of thoracic organs poses a severe challenge for the administration of image-guided radiotherapy treatments. Providing online and up-to-date volumetric information during free-breathing can improve target tracking, ultimately increasing treatment efficiency and reducing toxicity to surrounding healthy tissue. In this work, a novel population-based generative network is proposed to address the problem of 3D target location prediction from 2D image-based surrogates during radiotherapy, thus enabling out-of-plane tracking of treatment targets using images acquired in real-time. *Methods* The proposed model is trained to simultaneously create a low-dimensional manifold representation of 3D non-rigid deformations and to predict, ahead of time, the motion of the treatment target. The predictive capabilities of the model allow correcting target location errors that can arise due to system latency, using only a baseline volume of the patient anatomy. Importantly, the method does not require supervised information such as ground truth registration fields, organ segmentation, or anatomical landmarks. *Results* The proposed architecture was evaluated on both free-breathing 4D MRI and ultrasound datasets. Potential challenges present in a realistic therapy, like different acquisition protocols, were taken into account by using an independent hold-out test set. Our approach enables 3D target tracking from single-view slices with a mean landmark error of 1.8 mm, 2.4 mm, and 5.2 mm in volunteer MRI, patient MRI, and US datasets, respectively, without requiring any prior subject-specific 4D acquisition. *Conclusions* This model presents several advantages over state-of-the-art approaches. Namely, it benefits from an explainable latent space with explicit respiratory phase discrimination. Thanks to the strong generalization capabilities of neural networks, it does not require establishing inter-subject correspondences. Once trained, it can be quickly deployed with an inference time of only 8 ms. The results show the capability of the network to predict future anatomical changes and track tumors in real-time, yielding statistically significant improvements over related methods.

**Keywords:** Motion tracking, Deep generative networks, 4D MRI, 4D ultrasound, Radiotherapy, Liver

### 6.2 Introduction

Changes in organ shape and movement in abdominal and thoracic cavities due to the patient breathing induced deformation represent an important barrier in radiation therapy. Consequently, target tracking strategies are crucial to improve control of radiation beams within the body [202]. Several studies, particularly in hepatic imaging, have shown the extent of liver motion and deformation during free breathing, as well as between consecutive inhale and exhale phases [203]. These studies demonstrate that deformation modes is much more complex than a simple caudal/cranial translation, and includes elastic deformation as well as rotation effects which might affect the dose administration towards a defined target.

In recent years, technological innovations such as MR-Linac systems have enabled the integration of high-strength MR scanners with linear accelerators into a single device, providing high-quality, real-time images of tumor targets while these are exposed to radiation beams. However, the acquisition is limited to select 2D slices, which does not capture out-of-plane motion, thereby compromising an accurate adaptation for 3D motion tracking. Therefore, the ideal solution for in-room treatment systems would yield continuous 3D information of both tumour and surrounding tissues location. Currently, the acquisition of volumetric images at a clinically acceptable spatio-temporal resolution is not available in commercial scanners due to physical and physiological constraints. Hence, the reconstruction of a full volume from partial observations (in-room cine slices) is highly desirable for on-table needs. Moreover, to avoid errors during dose delivery, the system latency must be considered, which is the time required for surrogate acquisition, target detection, beam shaping calculation and delivery.

Traditional solutions rely on non-rigid registration and deformable statistical models [90, 178, 179, 202, 204]. In [202], the authors extrapolated the 2D displacement fields estimated between orthogonal cine-MRI and the corresponding slices in a reference volume in the three anatomical directions. Then, the obtained volumetric motion field was used to warp the reference volume and estimate the deformed volume. However, this approach is difficult to adapt in real-time. Alternatively, many works are based on maximizing the correlation between the surrogate images with motion models. An important consideration is that the target position has to be predicted into the future so that the radiation gantry can react to the estimated target motion in a reasonable timeframe. These motion models can be subject-specific or population-based. In the first category, the motion information is extracted from patient-specific 4D data using deformable registration. Generally, a statistical model is computed afterwards using this motion information. For instance, in [90] Principal Component Analysis (PCA) was applied to parameterize the pre-treatment motion information. The weights of the eigenvectors were iteratively optimized until achieving the best alignment between a warped reference volume and the surrogate slices. A similar approach was adopted in [175], assuming a pre-operative reference volume along with single online 2D cine MRI slices. The authors proposed to refine the motion model using free-from deformations and used a data fidelity constraint to find the best match between the warped volume and the 2D image during the optimization. Inspired by a previous work [178], in [179] was proposed a model based on regions of interest to relate the 3D motion, derived from 4D-CT data, with surrogate CT slices. In the literature, results reported for these patient-specific models are often more accurate than for population-based methods. In a clinical scenario, its reliability depends, however, on the degree of patient-specific inter-fraction motion variations. In addition, the assumption of a pre-treatment 4D MRI acquisition and processing constitutes an important limitation since it is not available in many clinical scenarios with traditional Linac systems. In contrast, population-based models can be applied in unseen patients in the absence of any 4D image data. To construct them, datasets of different patients are combined to a single motion model with the advantage of capturing broader motion variability (see Figure 6.1). Moreover, the model generation can benefit from a progressive increase in the amount of data [203]. In several cross-population approaches proposed over the past years, the backbone is a PCA model driven with multiple types of surrogates [205–207] although some works based on manifold learning have also been reported [208]. In [206], anatomical landmarks in ultrasound images were used for driving a population model. Furthermore, the authors employed an artificial neural network with a single hidden layer for spatio-temporal prediction of the respiratory liver motion. In [207], non-linear regression was applied to find the correlation between arbitrary surrogate signals and the motion model parameters. Although the aforementioned population models have achieved good accuracy with errors below 3 mm, establishing inter-subject correspondences remains a challenging task due to anatomical variations and missing correspondences in the presence of pathological structures. It is also time-consuming, complex, and involves direct human interaction.

Recent advancements in deep learning have opened new opportunities to address the 4D reconstruction task given sufficiently large training datasets. Several attempts were made with the purpose to learn a joint mapping between partial views and prior 3D shapes [209, 210]. These works have paved the way to relate partial observations with high-dimensional data in deep frameworks. Also, the excellent generalization capabilities of neural networks enable learning over a population dataset and applying the knowledge to unseen subjects, which resembles traditional inter-subject motion models. Despite this, there are relatively few works on motion modeling and 4D tracking using deep networks. A related work in this field using conditional generative networks was presented by [180]. However, it requires patient-specific 4D data before treatment, which is a major limitation. Moreover, the model was validated with only 2 seen anatomies, each one with a few hundreds of temporal points.

We propose a novel population-based framework to address the problem of online predictive 3D motion tracking from navigation surrogates. Our model is trained in an unsupervised manner, to learn a compact manifold representation of a population of various 3D deformations from a generative network, which allows for out-of-plane target tracking using only in-plane real-time images. Furthermore, the model leverages sequences of past partial observations to recover deformation fields ahead of time, thereby allowing for system latency compensation. To the best of our knowledge, this is the first population-based model using deep generative networks applied to respiratory motion tracking for 4D MR or US imaging. Its prediction capability in unseen anatomies finds an ideal application in radiation treatments where subject-specific 4D MRI acquisitions are not available.

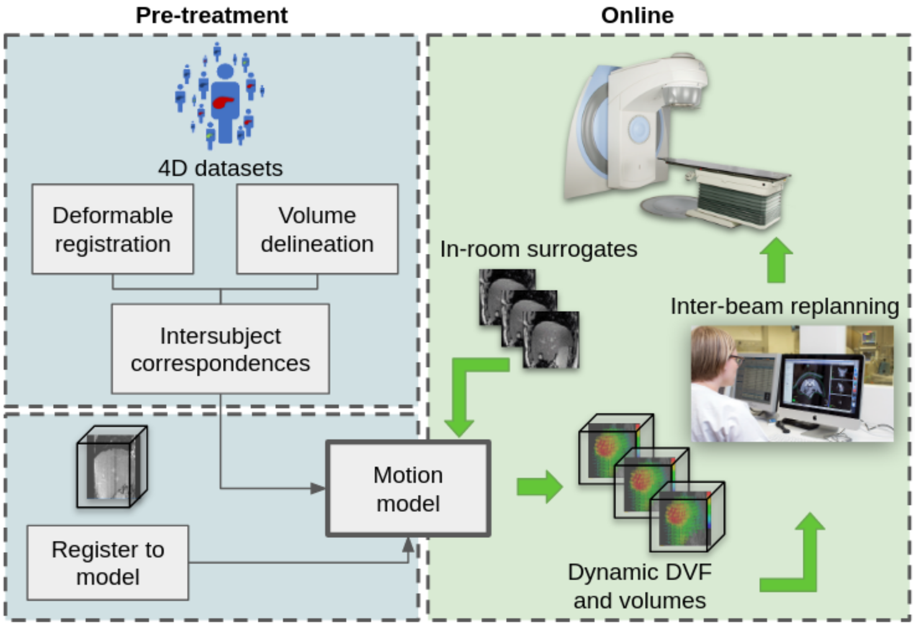


Figure 6.1 Schematic representation of the main components for population-based motion modeling before treatment and its application during image-guided radiotherapy.

## 6.3 Material and Methods

The proposed surrogate-driven deep motion model learns from population-based 4D datasets which were acquired and reconstructed to cover a significant variety of breathing patterns. Figure 6.2 shows a schematic representation of the proposed model. During training, the model learns, in an unsupervised way, how to map volume deformations at different respiratory phases to a low-dimensional space. Furthermore, it learns to recover the dense deformation given its compact representation. During the inference process, anatomical information is added in the form of compressed skip connections. Partial observations, specifically 2D images (cine MR or B-mode ultrasound), which contain phase information, are processed by a spatiotemporal network. Each phase representation is related to a certain deformation, represented by its corresponding low-dimensional coordinate. After training, the constructed embedding can be seen as a motion model which can be driven by the image surrogates. During deployment, the inputs are a pre-treatment volume gated at a common reference respiratory phase and 2D image surrogates. The model allows predicting a future 3D displacement vector field (DVF), thereby compensating for system latencies. Details about the dataset acquisitions as well as the model components are provided in the next subsections.

### 6.3.1 Datasets

**4D MRI:** Free-breathing sagittal slices were acquired from 25 healthy volunteers, who provided their written consent. The acquisitions were carried out on a 3T Philips Ingenia whole body MRI scanner using a 2D T2-weighted Balanced Turbo Field Echo (bTFE) sequence. This sequence allowed good vessel visualization without using any contrast agent. The acquisition was limited to blocks of 5 min each since longer acquisitions made the subjects feel uncomfortably warm. During sessions of 20 minutes, 4 blocks were acquired, which corresponded to 240-400 breathing cycles, considering that the normal respiration rates for an adult person at rest range from 12 to 16 breaths per minute. Image dimensions were $32 \times 176 \times 176$, pixel spacing was $1.7 \times 1.7$ mm$^2$ and the slice thickness was 3.5 mm. An alternation scheme was followed to acquire data frames covering the right liver lobe inter-

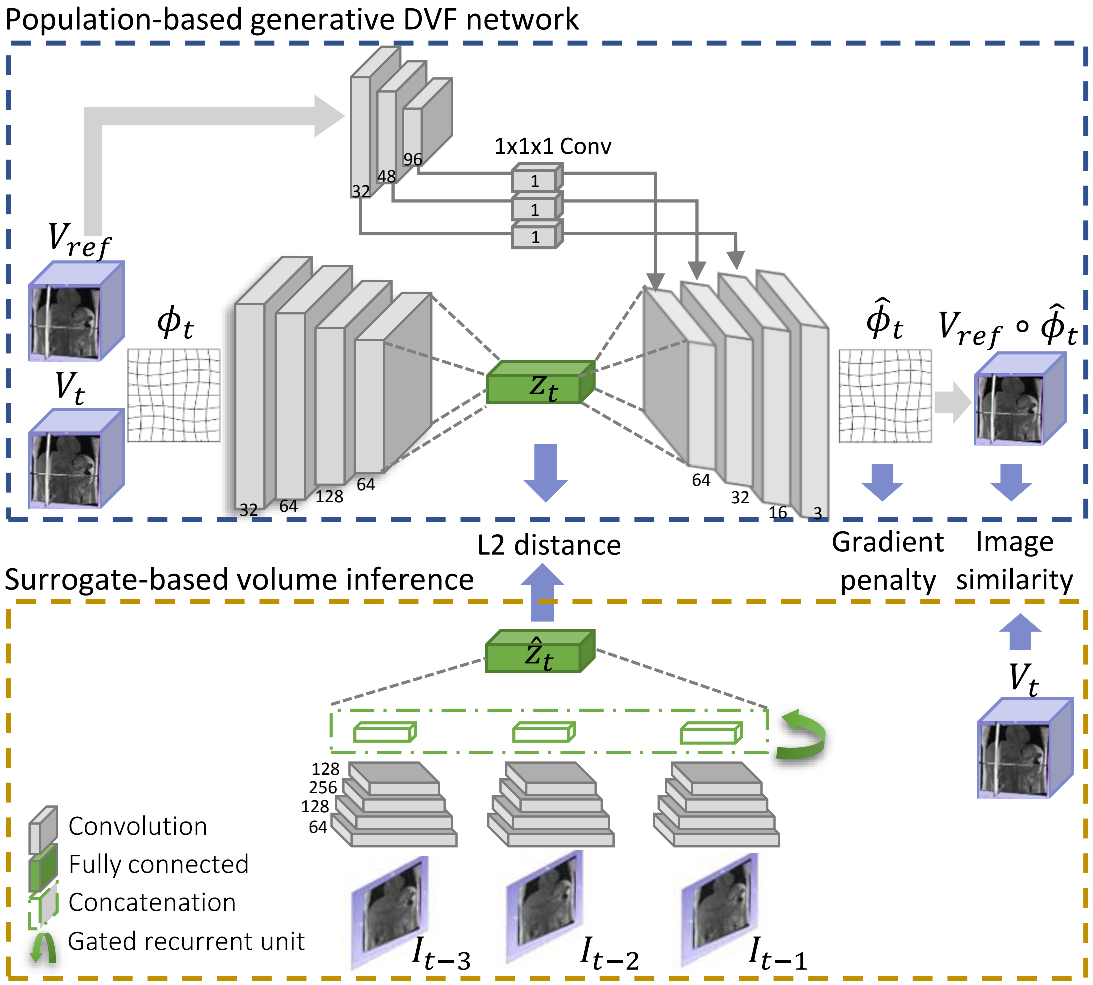


Figure 6.2 Proposed architecture for intra-treatment volume prediction from partial 2D observations and a static reference volume.

leaved with navigator frames taken at a fixed anatomical position, chosen in the middle of the liver. In order to produce time-resolved volumes, we followed the slice stacking approach detailed in [2]. The temporal resolution of the volumes was of 450ms, which produced 2480 reconstructed volumes per subject.

A second free-breathing MRI dataset was acquired in a 3.0 T clinical MRI system (Ingenia, Philips Healthcare) using a 3D stack-of-stars gradient-echo radial sequence with golden-angle sampling scheme. This sampling scheme, which uses $\approx 111.25°$ angular increment between consecutive spokes, enables extraction of the respiratory signal. Relevant imaging parameters included TR/TE=3.40/1.40 ms, flip angle 12°, field of view (FOV) $450 \times 450 \times 250$ mm$^3$, spatial resolution $1.5 \times 1.5 \times 5.0$ mm$^3$. The data was acquired continuously during 3 minutes and further reconstructed into 10 respiratory phases using the XD-GRASP technique [98]. The study population comprised 11 patients diagnosed with hepatocellular carcinoma, who provided their written consent. For each patient, tumors exceeding 10 mm in the right liver lobe were annotated by an experienced abdominal radiologist using previous diagnostic images. Of the patients, 6 were men and 5 women, with ages around 70 (± 11) years.

As a pre-processing step for model deployment, the volumes of both datasets were cropped to $32 \times 64 \times 64$ and resampled to a voxel size of $3.5 \times 3.4 \times 3.4$ mm$^3$ to focus on the liver and remove organs in the bottom part of the abdomen such as the stomach, pancreas, kidneys and intestines. Therefore, the modeled field of view was $\approx 112 \times 218 \times 218$ mm$^3$ in the left-right, anterior-posterior and superior-inferior anatomical planes, respectively. The 4D reconstruction from cine acquisitions is a challenging task that is not exempt from errors in the sorting process. Furthermore, because of some uncertainties during the image acquisition and the involved deformable registrations, the actual motion is not precisely known. Nonetheless, the 3D deformations used to train the model still represent a valid ground truth to evaluate its performance.

**4D US:** A third dataset of free-breathing 4D US sequences from 20 healthy volunteers, who provided their written consent, was acquired using a Philips EPIQ 7G ultrasound system with a X6-1 matrix array transducer. During acquisition, the ultrasound probe was placed under the sternum along the sagittal plane, capturing a cross section of the left liver lobe. The imaging depth was set to 12 cm. Focus and contrast were adjusted to provide the best visualization of the liver and its vessels. Limited to a 15 s acquisition window, it was possible to capture up to 3 respiratory cycles with a 250 ms temporal resolution. The acquired volumes were first pre-processed by applying a Bayesian non-local means filter [211] for speckle removal. Then, the volumes were resampled to a $2.0 \times 2.0$ mm$^2$ spatial resolution with a slice thickness of 1.0 mm and cropped to a volume size of $64 \times 64 \times 32$.

### 6.3.2 Problem formulation

We consider an ensemble of $P$ time-resolved 3D acquisitions over a population, generating 4D datasets. The motion observed in each dataset $p \in P$ with $T+1$ temporal volumes is described by deformation fields $\phi_t \in \mathcal{R}^{H \times W \times D \times 3}$ between a moving image $V_0 \in \mathcal{R}^{H \times W \times D}$ and the fixed images $V_t \in \mathcal{R}^{H \times W \times D}$ where $t \in [1, T]$, and $H, W$ and $D$ denote the height, width and depth of the volumes, respectively. For each subject, a volume obtained at exhale $V_{t=0}$ is selected, which serves as reference to determine relative displacements to each other temporal volume. Therefore, each dataset contains a set of 3D deformations $\phi^p = \left\langle \phi_1, \phi_2, \ldots, \phi_T \right\rangle$. The first goal is to compute a mapping between each deformation and its low-dimensional representation $\phi_t \rightarrow z_t \in \mathcal{R}^d$ where $d \ll H \times W \times D \times 3$, thereby computing a motion model. Additionally, it is assumed that for each deformation $\phi_t$, image surrogates at times $\Big\langle I_{t-1}$,$I_{t-2}$, $\ldots\ I_{t-m} \Big\rangle$ are available. Having this compact representation of motion, the second goal is to drive the model by using these partial observations.

### 6.3.3 Population-based generative DVF network

The first step of the workflow for motion modeling is motion quantization using deformable registration. In order to develop a fully differentiable pipeline, we use a registration function parameterized with a neural network. It receives a reference volume $V_{ref}$ and a target volume $V_t$ at time $t$ as inputs to generate the breathing-induced organ DVF $\phi_t$ between them. For the registration function, we use the U-net-like architecture proposed in [42]. Nevertheless, the proposed framework is agnostic to this module. Therefore, any other similar configuration can be used. The registration module is previously trained using the same training set, meaning that during model optimization their weights remain static.

Dimensionality reduction was shown to be an essential tool for motion modeling. The core idea is to uncover the structure of high dimensional data by projecting it down to a subspace where hidden features become visible. We leverage the capacity of autoencoders to learn a non-linear parametric mapping from volume deformations to their latent representations. The goal of the auto-encoding process is to produce a meaningful space at the bottleneck that enables input reconstruction. In our model, we start by defining a feature-extracting function denoted as $f_\theta$, namely the encoder. It computes a feature vector $z = f_\theta(\phi_t)$ from an input $\phi_t$ which is a compact input representation. Another function, $g_\theta$, called the decoder, maps from the low-dimensional representation back into the input space, thereby yielding a reconstruction $\hat{\phi}_t = g_\theta(z)$. The assumption is that by compressing the data into a more compact representation space, the model decides which features of the observed data are relevant information and what aspects can be discarded.

The proposed deep motion model is composed of a motion encoder, an auxiliary encoder for the anatomical information and a motion decoder. The first and second sub-models possess similar configurations except for the number of input channels (3 for the motion encoder and 1 for the auxiliary encoder) and output channels. They are composed of successive 3D convolutions with kernel size 3 × 3 × 3 and a stride of 2 followed by ReLU activations and batch normalization. Due to the ill-posedness of the autoencoding framework, we integrated an auxiliary encoder. Its role is to extract features from the reference volume. Since during model deployment only partial observations will be available, the pre-operative volume will be the only acquisition that will provide complete subject-specific anatomical information. Features go through compressed skip connections before reaching the decoder as a way to limit information bypass while preserving spatial detail [212]. In this variant, feature maps are compressed via 1 × 1 × 1 convolutions to a single map before concatenation in the decoder. Also, the features skipped from the anatomical encoder are normalized with instance normalization. The decoder receives a 256 sized latent vector $z$ ($z = 256$) which is reshaped and fed to a stack of convolutional layers followed by Leaky ReLU (0.2) non-linearities and batch normalization. The last convolutional layer has linear activation to output 3 feature maps corresponding to the motion planes. We train the autoencoder with an image similarity loss on the final voxel output against the target voxels ($V_t$) while ensuring smooth motion fields. This loss function has the form:

$$\mathcal{L}_{rec} = \mathcal{L}_{sim}\left(V_t, \hat{V}_t\right) + \alpha\mathcal{L}_{smooth}\left(\hat{\phi}_k\right) \tag{6.1}$$

where $\hat{V}_t$ results from warping $V_{ref}$ with the estimated motion $\hat{\phi}_k$ and $\alpha$ is a regularization parameter.

### 6.3.4 Surrogate-based volume inference

The lower part of the model illustrated in Figure 6.2 constitutes a module designed for driving the model for future volume inference. It receives an image sequence in $m$ prior time points with respect to the deformation in the deep model, ie. $\left\langle I_{t-1}, I_{t-2}, \ldots\ I_{t-m}\right\rangle$. Each single image is fed to a stack of 2D convolutional layers with kernel size 3 × 3 and a stride of 2 followed by ReLU activations and batch normalization. The feature vectors are concatenated through the temporal dimension and fed to a convolutional gated recurrent unit which leverages the spatiotemporal information. This is followed by a fully connected layer that produces an embedding vector $\hat{z}_t$, which attempts to resemble the encoded deformation $z_t$ by minimizing the $L_2$ distance between both latent representations. The optimization problem in our framework can be written as:

$$\arg\min_{\mathbf{\Phi}} \mathcal{L}_{total}\left(V_t, \hat{V}_t, \hat{\Phi}_t, z_t, \hat{z}_t\right) \quad \text{where} \quad \mathcal{L}_{total} = \mathcal{L}_{rec} + \beta \mathcal{L}_2\left(z_t, \hat{z}_t\right) \tag{6.2}$$

where $\beta$ is a regularization parameter.

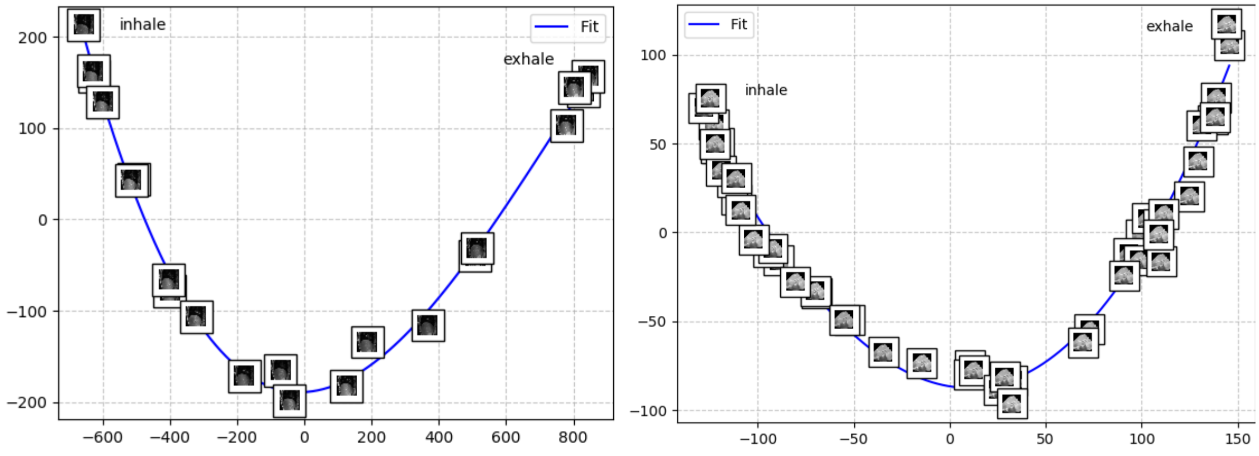
### 6.3.5 Implementation details

The proposed model has 2 main tasks: (1) to create a low-dimensional mapping by compressing and recovering 3D deformations, and (2) to associate partial observations to their corresponding embedding points. Training the proposed model from scratch to jointly address both tasks is a challenging problem. Therefore, the model was trained in three stages thereby learning both aforementioned tasks in the first two stages. In the first stage, we train the autoencoder independently to create the motion model using Eq. (6.1). The second stage focuses on regressing the latent code generated by the encoder from the input deformation field by solely minimizing $\mathcal{L}_2$. Three temporal points for the image surrogates were used ($m = 3$). Meanwhile, the autoencoder weights remain fixed. In the final stage, we fine-tune the network jointly with both losses but weighting the $\mathcal{L}_2$ term with a parameter $\beta = 0.01$ as shown in Eq. (6.2). In all steps, the network's parameters were optimized using the Adam optimizer with an initial learning rate set at $10^{-3}$, which was reduced by a factor of 2 after multiples epochs without improvement. The MRI dataset was trained using the negative local normalized cross correlation (NCC) while the US dataset was trained using Mean Square Error (MSE) as the image similarity metric. In Eq. (6.2), $\alpha = 0.01$ both for NCC and MSE. Mean centering and standard deviation normalization were applied to the input images and volumes. Training was performed in PyTorch with a batch size of 10. We used a leave-one-out validation scheme for both volunteers MRI and US datasets, considering a different anatomical case for testing. The patient MRI dataset was used for evaluation purposes as an independent hold-out test dataset. This means that the images were not used in any way during the model optimization. Our code is available at *https://github.com/lisetvr/population-TL-model.*

## 6.4 Experimental results and discussion

We evaluate the effect of the surrogate slice plane on the estimation accuracy. Sagittal and coronal plane images were considered as the surrogate navigator since they both capture the cranial-caudal direction where the largest liver motion occurs. Furthermore, the predictive capability of the proposed model was confronted to a motion extrapolation (ME) approach proposed by [202] in the context of MRI-guided radiotherapy. We also implemented a related deep network (DN) which fuses feature representations from a reference volume and a surrogate slice to generate a 3D deformation [213]. Both approaches are considered population-based, meaning that a subject-specific dataset is not required prior to treatment. Statistical significance was calculated by applying a Wilcoxon signed-rank test. In all the tests, $p < 0.01$ was considered to reject or fail to reject the hypothesis that the compared samples come from the same distribution.

In the first experiment, we investigated the structure of the latent space of the proposed model. We applied PCA on the latent code vectors to reduce their dimensionality to a single point in a bidimensional Cartesian space while retaining $\approx 92$ % of the variance. The manifolds shown in Figure 6.3 reveal that, for both modalities, data points are clustered according to their position within the respiratory cycle, which is convenient for a motion model. Indeed, the size of the latent vector $z$ is an important choice for the network design since it defines how much variability can be encoded in the model. Experiments that support this decision can be found in the supplementary materials. Our second experiment aimed at investigating how the model copes with inter-cycle variability and irregular breathing. A single liver vessel, located in the medial position and near to the diaphragm, was tracked through several respiratory cycles. Figure 6.4 illustrates the target and predicted relative vessel positions from MRI as well as the error plot in the superior-inferior and anterior posterior motion directions in three cases with irregular breathing. From the graph we can see that errors in the superior-inferior and anterior-posterior motion planes remain lower than 3 mm and 1 mm, respectively. In these cases, our model showed a reasonable performance



(a) MRI dataset
(b) US dataset

Figure 6.3 Low-dimensional mapping of the latent representation in MRI and US datasets.

following the target trajectory in the presence of small apneas and variable cycle amplitudes.

In order to assess the accuracy over the whole anatomy, 3D deformable registration between ground-truth and predicted volumes was performed using a B-spline transformation model

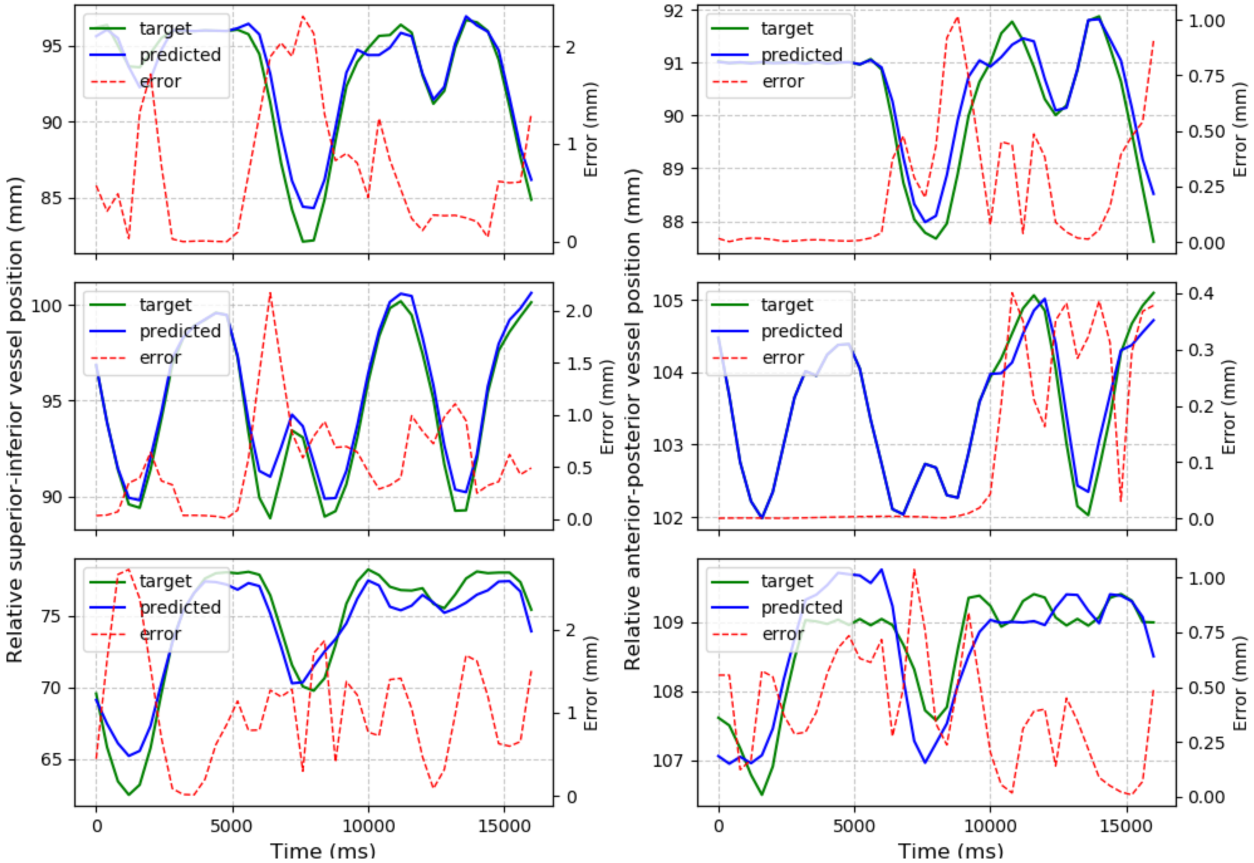
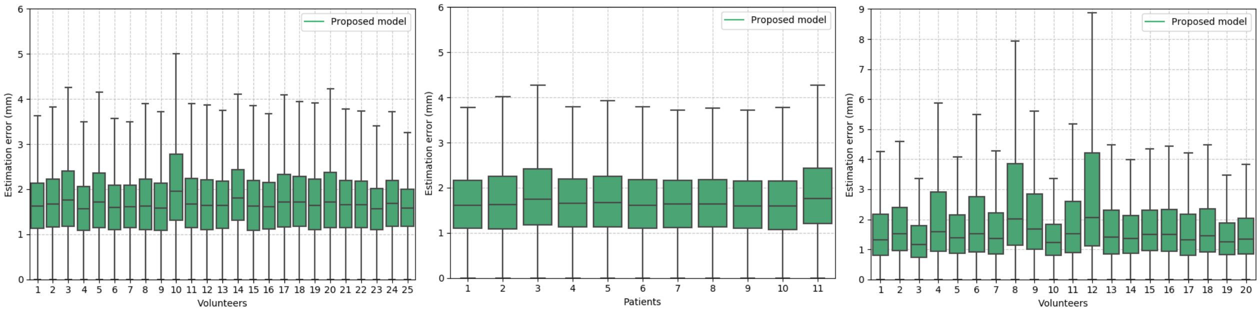


Figure 6.4 Vessel trajectories in the superior inferior and anterior posterior motion planes observed in three subjects with irregular breathing in the MRI dataset. Dashed red lines represent the error (in mm).



(a) MRI volunteer dataset

(b) MRI patient dataset

(c) US dataset

Figure 6.5 Spatio-temporal prediction errors of all voxel-wise displacements.

and the mutual information similarity measure implemented in Elastix [214]. This toolbox is widely used for linear and non-linear registration of abdominal images. Organ masking is commonly used when matching lung data to avoid considering the rib cage, especially in CT scans. We considered the whole MR volume since the lung area is minimal and the rib cage is not visible. Moreover, relatively small displacements are expected since, overall, predicted volumes are similar to ground-truth volumes. Figure 6.5 presents the error distribution considering all spatio-temporal voxel-wise displacements. It can be observed that the median error is smaller than 3 mm across all the datasets. The overall computed mean error was 1.78 ± 1.0 mm, 1.74 ± 0.9 mm and 1.99 ± 1.98 mm for volunteer MRI, patient MRI and US datasets, respectively. Between 5 and 10 expert-selected landmark annotations throughout one complete respiratory cycle were used to measure the geometrical accuracy between ground-truth and predicted landmark positions (see Figure 6.6). During inference, we excluded the reference volume from the processed volumes. Therefore, it was not considered in the calculation of the prediction errors. Moreover, the reference volume was taken at the very first end-exhale obtained, while the evaluated breathing cycle was the last one in the dataset. Hence, the elapsed time between both was maximized within the limits of the dataset (15-20 min interval for MRI). Tables 6.1 and 6.2 summarize the target tracking errors computed on both datasets for different respiratory phases. For reference, in the first row of the tables, we report the tracking errors measured when there is no motion compensation (Unregistered). The values reveal that using coronal plane slices yield an increased performance compared to the sagittal view, presumably because the coronal plane covers a larger liver area. Moreover, we found the differences between measurements obtained using sagittal and coronal slices as being statistically significant ($p < 0.01$, Pearson correlation coefficient $\rho = 0.94$). In the MRI dataset, our model driven by coronal slices has demonstrated the ability to accurately predict deformations throughout all the respiratory cycle. The most challenging predictions were near the inhale phase, where the registration-based approach led to the lowest median errors. It should be noted this phase is prone to inter-cycle variability. Furthermore, large displacements occur due to increased volume of the lungs during air intake. When comparing overall median results, the accuracy improved by a statistically significant margin by 1.7 mm ($p < 0.01$, $\rho = 0.92$) and 1.0 mm ($p < 0.01$, $\rho = 0.82$) over DN and ME approaches, respectively. In the US dataset, the proposed model outperformed both DN and ME, achieving a statistically significant improvement of 1.7 mm ($p < 0.01$, $\rho = 0.91$) and 0.9 mm ($p < 0.01$, $\rho = 0.82$) overall respectively. Generally, both slice orientations provided similar performances except for the inhale phase where the coronal surrogate model achieved a lower average TRE. In the US dataset, the sagittal view covers a larger portion of the liver than the coronal view. We hypothesize that this contributes to the better performance of the model with the sagittal surrogate overall. Figure 6.7 displays the variability over subjects across all the evaluated datasets.

Figure 6.8a presents NCC values between ground-truth and predicted volumes when the imaging plane is shifted from the middle liver position, with which the model was trained. It can be seen that, in both MRI and US datasets, the similarity values remain approximately constant. Therefore, the model is tolerant to potential shifting of the surrogate plane. This characteristic is especially important for ultrasound, where it is more difficult to reproduce a certain imaging plane. We also compared predicted and ground-truth volumes at 5 different sub-volumes along the right-left axis to evaluate if the model better predicts the motion in the vicinity of the surrogate plane across all the MRI and US datasets. Figure 6.8b shows that, in the MRI dataset, the quality is slightly degraded on the left side of the volumes. However, in this case the leftmost slices correspond mainly to skin and ribs. On the other hand, in the US dataset there is a relatively stable similarity across all positions.

Difference maps of the temporal volumes at several respiratory phases in MRI and US datasets are shown in Figures 6.9 and 6.10. In both cases, comparing ground truth and predictions, it is noticeable that the model correctly predicts the motion shown by the true image sequence. Additional qualitative results can be found in the supplementary materials. Finally, we assess the plausibility of the deformations by computing the Jacobian matrix determinant ($|J|$). The percentage of voxels with a non-negative $|J|$ was 99.7%.

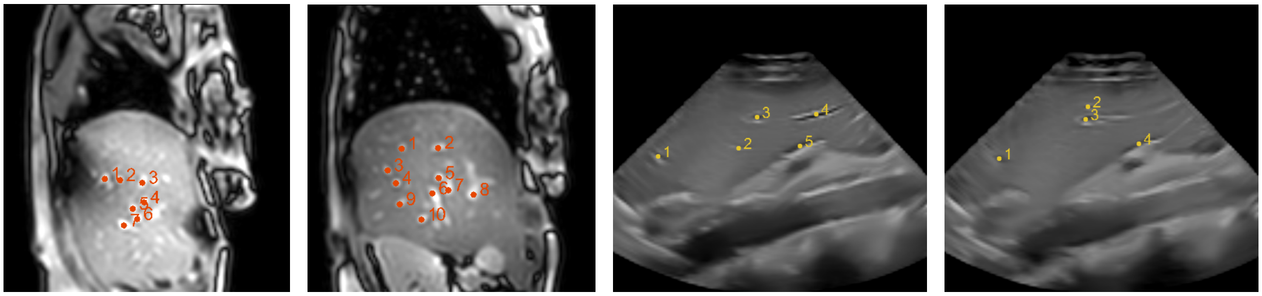
The proposed method requires a mean computation time of 8 ms (average from 20 measurements) for predicting the deformation field on a NVIDIA Titan RTX GPU with 64 GB RAM. With a prediction horizon of 450 ms (in the MRI dataset), the motion model is real-time applicable and allows for online tracking of the target volume. Typical system latencies encountered during dynamic target tracking based radiation delivery are estimated to be of the order of 300 ms [203]. The capacity of convolutional recurrent units for image-based sequential prediction has been previously validated in [119] for MRI and US imaging modalities. In this work, we applied this sort of structure to extrapolate one future time step, which depends on the temporal resolution of the employed acquisition. Nonetheless, the predictive horizon can be extended to more time steps. This should be validated in a future study. After comparing results achieved in both datasets, it is noticeable that the performance in the US dataset was worse than in the MRI dataset. It is important to note that the former captured less respiratory cycles and by consequence, less motion variability. Also, the US dataset is comprised of a smaller number of subjects. It is well known that, in deep learning-based approaches, the dataset size is a limiting factor, particularly given the poorer image quality. Therefore, to create robust deep motion models, it is crucial to capture enough breathing

Table 6.1 Target tracking errors (in mm) measured at selected respiratory phases for the model trained with the MRI dataset. Overall values consider all the phases. Values are mean, median, ($P_{95}$).

| **Model** | **Mid-inhale** | **Inhale** | **Mid-exhale** | **Exhale** | **Overall** |
|---|---|---|---|---|---|
| Unregistered | 9.5<br>8.4 (15.6) | 11.6<br>10.8 (18.9) | 4.4<br>2.5 (12.1) | 1.6<br>0.3 (8.5) | 6.6<br>5.5 (17.4) |
| DN [213] | 4.8<br>4.5 (12.4) | 5.3<br>2.9 (10.3) | 3.1<br>2.1 (5.5) | 2.6<br>1.3 (4.7) | 4.1<br>3.1 (11.4) |
| ME [202] | 3.2<br>2.8 (8.8) | **2.5**<br>**1.7 (6.1)** | 2.2<br>2.1 (4.6) | 2.3<br>1.5 (4.0) | 3.9<br>2.4 (8.2) |
| Proposed (sag) | 2.8<br>2.1 (6.1) | 4.9<br>3.0 (7.8) | 1.8<br>1.4 (4.2) | 2.3<br>1.2 (4.8) | 2.6<br>2.0 (7.1) |
| **Proposed (cor)** | **2.4**<br>**1.8 (6.0)** | 2.3<br>1.9 (4.0) | **1.3**<br>**1.1 (2.7)** | **1.0**<br>**0.8 (1.9)** | **1.8**<br>**1.4 (4.7)** |

Table 6.2 Target tracking errors (in mm) measured at selected respiratory phases for the model trained with the US dataset. Overall values consider all the phases. Values are mean, median, ($P_{95}$).

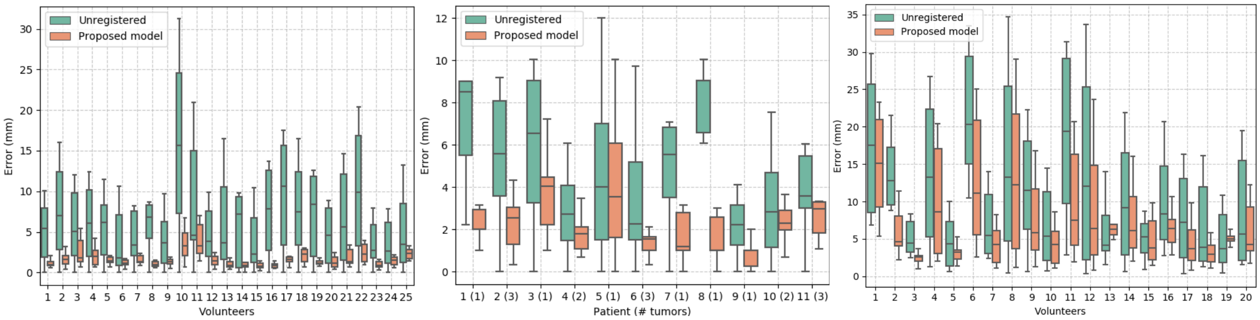
| **Model** | **Mid-inhale** | **Inhale** | **Mid-exhale** | **Exhale** | **Overall** |
|---|---|---|---|---|---|
| Unregistered | 12.3<br>10.7 (24.8) | 17.9<br>17.2 (23.7) | 7.4<br>5.9 (14.3) | 4.1<br>2.8 (10.1) | 10.4<br>9.4 (26.9) |
| DN [213] | 10.3<br>9.3 (22.4) | 14.5<br>13.3 (24.6) | 6.1<br>4.6 (11.2) | 4.4<br>3.1 (9.1) | 8.8<br>6.9 (23.6) |
| ME [202] | 8.6<br>7.5 (17.8) | 12.9<br>11.9 (24.5) | 6.2<br>5.5 (12.0) | 4.1<br>3.4 (7.8) | 8.0<br>6.1 (22.4) |
| Proposed (sag) | 7.5<br>7.3 (13.7) | 12.1<br>12.7 (20.5) | **4.8**<br>**4.0 (8.1)** | **3.2**<br>**2.9 (7.1)** | **6.9**<br>**5.2 (17.0)** |
| **Proposed (cor)** | **7.9**<br>**6.2 (14.7)** | **11.5**<br>**10.2 (23.7)** | 4.9<br>4.6 (10.3) | 4.0<br>3.1 (8.6) | 7.1<br>5.4 (20.1) |



(a) (b) (c) (d)

Figure 6.6 Selected vessels across temporal volumes: (a) V5 MRI (medial), (b) V5 MRI (lateral), (c) V10 US (medial) and (d) V10 US (lateral).

variability and different anatomies. The main limitation of our motion model seems to be reaching the deformation at the inhale phase. Although the registration-based approach can be more accurate for this particular case, it relies on two orthogonal planes whereas our model uses only one. Furthermore, previous methods are limited to derive the global anatomy. Also, the run-time to acquire and register two pairs of images might be orders of magnitude greater than our approach, which is a major limitation for real-time interventions. The proposed framework is capable of predicting not only the 3D tumor position but the whole anatomy with a single imaging plane. Therefore, in the context of image-guided radiotherapy (IGRT), this knowledge can be used to estimate the delivered dose and subsequently to adapt the treatment. Besides, our approach presents two main advantages over those using 2 orthogonal images. First, the acquisition time of the surrogate is lower since we only use a single image rather than two. Second, we can track organs and structures that are not present in the imaging plane. In a prospective clinical study about the usage of IGRT to treat abdominal malignancies, it was found that, while the use of 2D-cine gating

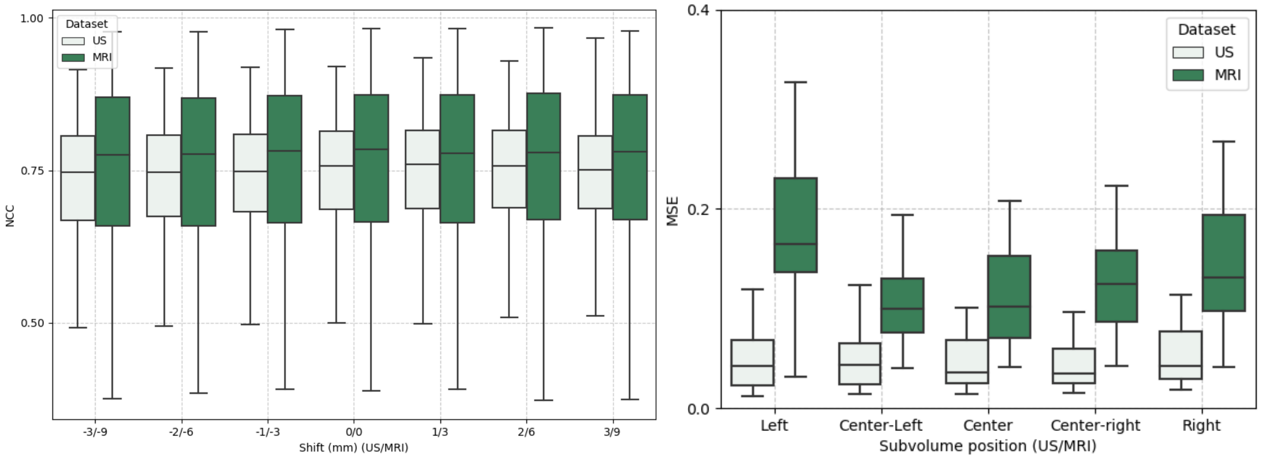


(a) MRI volunteer dataset (b) MRI patient dataset (c) US dataset

Figure 6.7 Initial motion and target tracking errors for each subject.

permitted target monitoring during treatment, the unobserved intra-fraction organs at risk (OAR) motion degrades dosimetric benefits [215]. It was also acknowledged that this concern could be mitigated by proving volumetric information. However, this real-time volumetric information cannot be provided by current scanners. Hence, so far, respiratory motion models remain the only available solution to enable real-time 3D tumor and anatomy tracking in combination with real-time online plan adaptation. On the other hand, a recent study on current challenges in IGRT has acknowledged that deep learning might play an important role towards the widespread clinical use of MR-guided radiotherapy [216]. For instance, in terms of pseudo-CT generation, automatic contour suggestion and deep motion models, such as the one we proposed. All of this showcases the need for future developments and streamlining of the motion model construction. The compared deep network (DN), which fuses features from a reference volume and surrogates, yielded the poorest performance. This indicates that our model benefits from a structured latent space, as previously illustrated in Figure 6.3, and from the addition of prior anatomical information during the decoding stage. The manifold representation also fosters the model's interpretability.

In the literature, lower errors have been reported using statistical modeling. Many of these works have limited the prediction to the right liver lobe [13, 182, 184, 206, 207]. In this work, the limited FOV was due to the acquired data. The MRI dataset used for model creation was originally acquired with a FOV covering only the right liver lobe. In contrast, the hold-



(a) (b)

Figure 6.8 Image similarity between ground-truth and predicted volumes. (a) Experiment where the location of the surrogate plane is shifted up to 3 mm and 9 mm in both directions from the middle position in US and MRI datasets, respectively. (b) Image quality at 5 different anatomical positions along the right-left axis in US and MRI datasets.

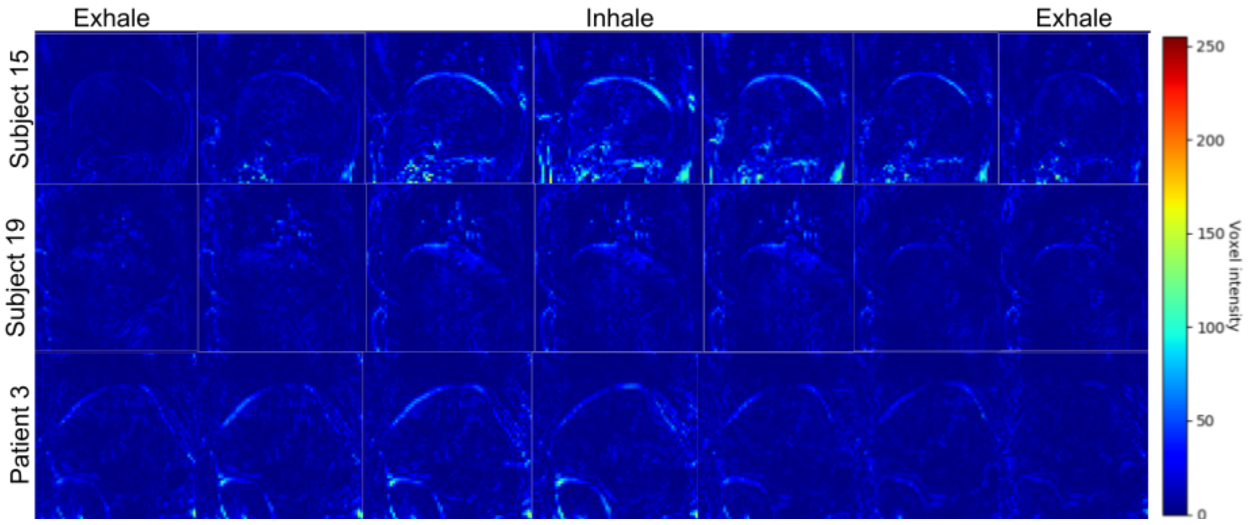


Figure 6.9 Difference maps between ground-truth and predicted volumes for two volunteers and one patient in the MRI datasets.

out test set was acquired with a larger FOV, imaging the whole abdominothoracic area. Therefore, prior to the model inference, the volumes were cropped to look similar to the ones used during training. Nevertheless, it should be noted that having a larger FOV is important for radiotherapy-related applications in order to perform dose calculations and organs at risk monitoring. In this case, our approach is still applicable since its working principle is independent of this feature. Although this would require more memory, modern GPUs should be capable of coping with larger FOV, as shown in a related work [180].

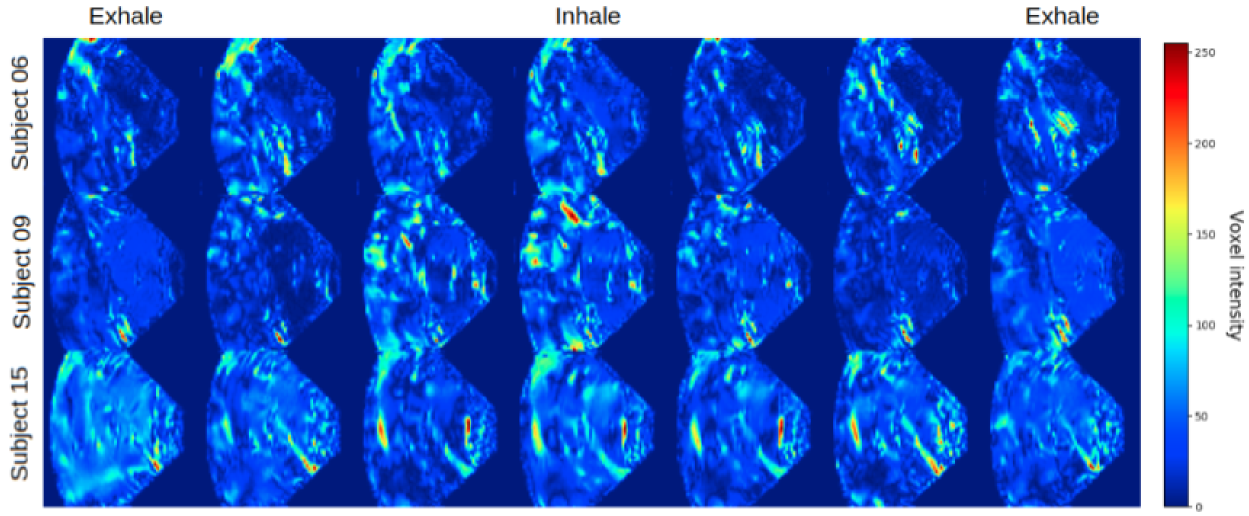


Figure 6.10 Difference maps between ground-truth and predicted volumes for three cases in the US dataset.

When comparing shape-based and landmark-based approaches for establishing inter-subject correspondences, Tanner, Yang, Samei and Székely [217] found that the former approach led to the lowest errors and enabled the motion prediction of the whole liver with a $95^{th}$ percentile of the errors below 5 mm. However, whether landmark or shape-based, establishing inter-subject correspondences remains an important limitation.

This process often requires manual intervention and can be complex and time-consuming, especially in large datasets. Our method relies on the strong generalization capability of deep networks to find patterns across a population dataset. In other words, the step equivalent to finding inter-subject correspondences in classical models is replaced by the unsupervised feature learning performed by our framework across the population samples. In our opinion, this represents a significant benefit over the state-of-the-art. Nonetheless, it should be noticed that it is assumed that there will be a correspondence in terms of field of view between the training volumes and the volumes at inference time. For proton therapy, where accuracy is more critical than in radiation therapy, a 2.4 mm average accuracy has been achieved with a classical PCA-based population model assessed in 8 volunteers [206]. According to this work, clinically acceptable accuracy for motion prediction and compensation should be less than 3 mm, which is achieved in our work. In summary, our population-based model can be deployed without any prior annotation steps while maintaining an acceptable accuracy. Although the model has been treated as population-based and hence validated on unseen cases, it can be readily adapted to work on subject-specific conditions. In this case, the model could be personalized (via fine-tuning) using temporal samples from the patient. This would lead to a better fit to the patient's needs and hence an increased accuracy. With the progressive expansion of the MR-Linac in the radiotherapy units and the promising results shown by fast 4D reconstruction strategies [64, 218], having a 4D MRI dataset for model personalization prior to treatment could be a viable option bringing important advantages.

Finally, as a proof of concept the 3D MRI reference volume used in the reported experiments was extracted from the 4D-MRI dataset. However, in the clinical scenario this reference volume will be a breath-hold image acquired before therapy. From a theoretical point of view, we hypothesize that this difference should not represent an obstacle since the model creates a low-dimensional representation regardless of the appearance of the reference volume. In other words, the embeddings are created from the displacement fields delivered by the previous registration module. This should be validated in a future study.

## 6.5 Conclusion

In this work, we presented a novel predictive population-based framework for real-time 3D motion tracking from 2D image surrogates. Our model is able to predict and generate accurate deformation fields with a temporal advance which allows for system latency compensation during radiotherapy treatments. Our model has shown promising results on two modalities, namely on MRI and US. The presented experiments have shown that our model is able to outperform comparative methods using only 1 imaging plane as a surrogate, while providing clinically acceptable target tracking accuracy under 8ms. Our model also does not require any prior processing steps such as surface segmentation or inter-subject correspondence identification. Future studies will assess the model robustness with regards to inter-fractional variations as well as the impact of the motion mitigation in the dose delivery.

**Funding:** This work was partly funded by an NSERC collaborative research and development project (CRDPJ-517413-17) and by the Canada First Research Excellence Fund through the TransMedTech Institute.

**Conflict of interest:** The authors declare that they have no conflict of interests.

**Ethical approval:** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Informed consent:** Informed consent was obtained from all individual participants included in the study.

# CHAPTER 7 ARTICLE 3: PROBABILISTIC 4D PREDICTIVE MODEL FROM IN-ROOM SURROGATES USING CONDITIONAL GENERATIVE NETWORKS FOR IMAGE-GUIDED RADIOTHERAPY

Contribution of the first author in preparation and writing this paper is evaluated as 90%. This article has been published by Medical Image Analysis journal on September 2021.

**Remarks:** This paper presents a probabilistic deep motion model, which can be deployed both as population-based or as subject-specific. The proposed model employs a conditional variational autoencoder as backbone to establish correspondences between respiratory phases and dense motion fields. Furthermore, it can generate multiple future volumes in one shot. In the test stage, it only requires a static 3D volume and cine 2D slices to predict future deformations. Experiments revealed that this approach yields a clinically relevant accuracy while presenting important advantages over similar state-of-the-art methods.

**Probabilistic 4D predictive model from in-room surrogates using conditional generative networks for image-guided radiotherapy**

Liset Vázquez Romaguera[1], Tal Mezheritsky[1], Rihab Mansour[2], Jean-François Carrier[3], Samuel Kadoury[1,2]

[1] Polytechnique Montreal, Canada, [2] Centre Hospitalier de l'Université de Montréal Research Center, [3] Centre Hospitalier de l'Université de Montréal and Département de physique, Université de Montréal, Montréal, Canada

## 7.1 Abstract

Shape and location organ variability induced by respiration constitutes one of the main challenges during dose delivery in radiotherapy. Providing up-to-date volumetric information during treatment can improve tumor tracking, thereby increasing treatment efficiency and reducing damage to healthy tissue. We propose a novel probabilistic model to address the problem of volumetric estimation with scalable predictive horizon from image-based surrogates during radiotherapy treatments, thus enabling out-of-plane tracking of targets. This problem is formulated as a conditional learning task, where the predictive variables are the 2D surrogate images and a pre-operative static 3D volume. The model learns a distribution of realistic motion fields over a population dataset. Simultaneously, a *seq-2-seq* inspired temporal mechanism acts over the surrogate images yielding extrapolated-in-time representations. The phase-specific motion distributions are associated with the predicted temporal representations, allowing the recovery of dense organ deformation in multiple times. Due to its generative nature, this model enables uncertainty estimations by sampling the latent space multiple times. Furthermore, it can be readily personalized to a new subject via fine-tuning, and does not require inter-subject correspondences. The proposed model was evaluated on free-breathing 4D MRI and ultrasound datasets from 25 healthy volunteers, as well as on 11 cancer patients. A navigator-based data augmentation strategy was used during the slice reordering process to increase model robustness against inter-cycle variability. The patient data was used as a hold-out test set. Our approach yields volumetric prediction from image surrogates with a mean error of 1.67 ± 1.68 mm and 2.17 ± 0.82 mm in unseen cases of the patient MRI and US datasets, respectively. Moreover, model personalization yields a mean landmark error of 1.4 ± 1.1 mm compared to ground truth annotations in the volunteer MRI dataset, with statistically significant improvements over state-of-the-art.

**Keywords** Motion modeling, Liver, Conditional generative networks, Radiotherapy, 4D imaging, Temporal prediction

## 7.2 Introduction

Radiation therapy is a well-established modality to treat malignancies in the thoracic and abdominal regions. This treatment modality uses ionizing radiation to destroy tumor cells. Its goal is to deliver the prescribed dose to the tumors while sparing healthy tissues and nearby organs [4]. However, organ shape and location variability induced by the patient's respiration during free breathing represents one of the main challenges during dose delivery. For instance, organs like lungs, liver, kidneys, and bowel, among others, are subject to respiratory motion, which have a large dosimetric impact, thereby compromising the treatment's effectiveness. Previous studies have shown that the total organ motion seen during treatment is composed of a main quasi-periodic component and other modes of deformation caused by the cardiac motion, digestive activity and muscle relaxation [219]. Moreover, even for the same patient, variations in breathing depth and speed over time may occur. Hence, although breathing shows a repetitive pattern, there is an inter-cycle variability that is non negligible.

Existing solutions for interventional organ motion management can be classified into two categories: non-adaptive and real-time adaptive methods. Within the first category, a simple method consists of asking the patient to interrupt breathing while the radiation dose is being delivered. Alternatively, the radiation beam can be turned on only during a certain period of the breathing cycle, which is known as gating. Both approaches require reproducibility of the organ position for the selected breathing phase and increase the procedure time. Clinical guidelines as well as other non-adaptive techniques are available in [22].

Real-time adaptive tracking is another motion compensation category designed to re-position the radiation beam as the target moves. Therefore, the accuracy of the dose delivery depends on the system adapting to the moving target anatomy. The success of such adaptation is related to the typical time delay between detecting a change in target position and the system change. For example, linear accelerators require a certain amount of time for adaptation. During such time, the target continues to move, thereby causing a perennial lag in the system response with respect to the target position [203]. Predicting the target position in advance is an approach to ensure that the radiation beam encompasses the target as it moves throughout the respiratory cycle. Toward this end, organ motion modeling, whether local or global, and temporal predictive mechanisms are crucial components.

Local approaches use information surrounding the target to exclusively estimate the tumor position, while global approaches relying on in-room surrogates (correlated signal acquired during treatment) and respiratory motion models estimate the whole anatomy. In previous works, surrogates are also referred to as partial observations [49]. Generally, forecasting mechanisms act over the surrogates to meet temporal requirements. For instance, linear adaptive filtering [184], multi-layer perceptron [174, 206] and recurrent neural networks [119] have been proposed for this purpose.

Commercial systems, such as the CyberKnife (Accuray) or Vero (BrainLAB), use correspondence models to estimate the internal tumor position as a function of external surrogates. The information provided by these systems is limited to the tumor position, which is generally represented by its center of mass or other fiducial marker, thus ignoring the surrounding anatomy. Some studies revealed low correlations between external surrogates and the internal organ motion. Therefore, in the clinical routine, the external surrogate is combined with low-frequency kV imaging. This allows the training and update of the correlation models, while controlling the non-therapeutic ionizing dose with respect to high-frequency fluoroscopy [48].

Recent technological innovations have enabled the integration of linear accelerators (Linac) with high-quality imaging capabilities during treatment [220, 221]. For instance, an MR-Linac can acquire an MRI with the patient lying on the treatment table [53, 222]. This novel paradigm, known as MR-guided radiotherapy (MRgRT), enables non-invasive monitoring of moving structures with excellent soft-tissue contrast [223] without the extra burden of ionizing radiation for the patient [216]. Similarly, other treatment modalities benefit from image guidance. For instance, ultrasound (US) images are acquired for radiofrequency ablation. Furthermore, high-intensity focused ultrasound (HIFU), can be guided using MRI [224, 225] as well as diagnostic US data [226, 227]. Recent studies have suggested that image-guided radiotherapy (IGRT) systems may improve treatment accuracy and control the toxicity in the surrounding healthy tissue for moving targets [228, 229].

Image acquisition during treatment with MRgRT is limited to 2D cine slices, which does not capture out-of-plane motion. While this still can be useful for tumor trailing [230] and stereotactic radiation treatments [231] it is well-known that tumors in the abdominothoracic area may exhibit a 3D hysteresis trajectory [13]. For this reason, the ideal imaging strategy to guide treatment delivery should yield real-time volumetric information. Additionally, the knowledge of the 3D tumor position would facilitate the reconstruction of the dose delivered during the treatment fraction. This, in turn, can serve for adaptive planning process for the next treatment fraction [64, 232]. The core idea is to use the real-time cine MR images acquired during beam-on to yield synthetic 3D+t volumes at the temporal resolution of the surrogate images. In the next section, we present the state-of-the-art on surrogate-driven motion modeling in the context of radiotherapy applications, with particular emphasis on MRgRT.

### 7.2.1 Related works

Current solutions deriving volumetric information are based either on non-rigid registration [202, 204, 233] and deformable statistical models [48, 90, 174, 175, 178, 179]. The former approach, also denominated as *fast/simplified strategy*, employs 2D-3D image registration between orthogonal cine-MRI and a pre-treatment reference volume to estimate the 3D target position. For instance, the method introduced by [204] completes the missing out-of-plane information with a phase-specific correction term computed a-priori. The computation of such correction term relies on a mask of the diaphragm of each sagittal slice of the 4D MRI data and optical flow registration. The method proposed by [202] is also purely based on image registration. The corresponding slices of a reference pre-treatment volume are registered to in-room sagittal and coronal cine-MRI. The 2D motion fields components are then replicated to all the slices in the volumetric space thereby yielding a 3D displacement vector field. A common limitation of these simple yet effective techniques is the inability to compensate for system latencies in real-time, with high computational requirements.

Alternatively, many works are based on maximizing the correlation between surrogate signals with organ motion models. These capture the motion and deformation of the internal anatomy due to respiration. The origin of the data used for its construction classifies them as subject-specific or population-based models. In both categories, the motion estimation from the dynamic volumes using deformable registration is a common step. In the first category, a statistical model is computed from the patient-specific motion information. These works rely on the maximization of a similarity metric between a surrogate image with the corresponding slice of a reference volume, which is iteratively warped until convergence is reached. [90] applied Principal Component Analysis (PCA) to parameterize the pre-treatment motion information. The weights of the eigenvectors were iteratively optimized until achieving the best alignment between a warped reference volume and the surrogate slices. Similarly, [175] proposed to refine the motion model using free-form deformations. Such an approach has been extensively validated using MRI and kV projections [173, 174, 176, 177]. In the same manner, [178] created a PCA model to establish correlations between 2D navigator images and 3D displacements. [179] revisited this concept with a region-based approach, enabling a local adaptation. In the literature, tracking errors reported for patient-specific models are often lower than those for inter-subject. Nevertheless, in the clinical context, its reliability depends on an accurate response to inter-fraction motion variations. Furthermore, due to time constraints, in many clinical scenarios it is not possible to acquire a patient-specific 4D dataset just before treatment. Conversely, population-based models constructed from different patients can be applied to new anatomies. Some authors argue that these models can benefit from large dataset to capture broader motion variability [203].

Similarly as in the subject-specific modeling, cross-population approaches have used statistical techniques to create a compact motion representation which is further driven with certain surrogates [13, 205–207]. Other works suggested the use of manifold learning theory for this task [208, 234]. [235] proposed the construction of an "exemplar model" by weighting the predictions of multiple subject-specific sub-models. In a later study, the exemplar model was driven using vessel landmarks which were tracked and temporally extrapolated via linear adaptive filtering [184]. [206] used anatomical landmarks in ultrasound images to drive a population model. Furthermore, they employed an artificial neural network with a single hidden layer for spatio-temporal prediction of the respiratory liver motion. [207] applied non-linear regression to find the correlation between arbitrary surrogate signals and the motion model parameters. Although the aforementioned population models have achieved good accuracy with errors below 3 mm, establishing inter-subject correspondences remains a challenging task due to anatomical variations and missing correspondences in the presence of pathological structures. It is also time-consuming, complex, and involves direct human interaction.

Recent advancements in deep learning have led to new opportunities for the image-based 3D object reconstruction task [236]. These data-driven approaches automatically discover and learn discriminatory features from image sets. Several attempts were made with the purpose of learning a joint mapping between partial views and 3D shapes [209, 210]. These works have paved the way to relate partial observations with high dimensional data in an end-to-end trained deep framework. Although 3D reconstruction from single 2D images has been an active area of research in the computer vision community, contributions in medical image analysis are rather limited to segmentation and shape reconstruction tasks [237–239]. Furthermore, most proposed architectures rely on annotated data such as triangulated meshes, binarized maps and point clouds and/or large datasets, which constitutes important limitations for clinical interpretation. Moreover, direct generation of grayscale intensities is challenging due to the wide range of values each voxel can take. In the motion modeling field, patient-specific and population motion models have been reported by [180] and [190]. In the former, a conditional generative adversarial network was developed to relate 2D US images with a 4D MRI dataset. This subject-specific model was validated with only 2 anatomies, each one with few hundreds of temporal points. In the latter case, the authors proposed to minimize the L2 distance between temporal low-dimensional feature representations and deformation encodings obtained with a convolutional autoencoder.

Other recent approaches have attempted to directly yield real-time volumetric images (MR-SIGMA) [99] or motion fields (MR-MOTUS) [100]. The key idea of MR-SIGMA is to match real-time acquired motion signatures with a set of pre-learned motion states. The main limitation is the adaptation to organ drifts and patient movement. On the other hand, MR-MOTUS recovered 3D motion fields directly from k-space data. Furthermore, it yielded plausible deformations with a predictive horizon of 170 ms on 5 subjects. Although this technique showed promising results, it requires further validation.

### 7.2.2 Contributions

As shown previously, population models using data-driven approaches are still very incipient. In this work, we propose a novel population-based deep probabilistic model to address the problem of real-time 3D motion compensation from image surrogates, with personalization capabilities when patient-specific data is available. It is one of the first population-based models using generative networks into the field of respiratory motion modeling and 4D MR imaging. The rationale behind this work is to leverage the representational capacity of conditional variational autoencoders (CVAE) [240] as a backbone to map organ deformations over

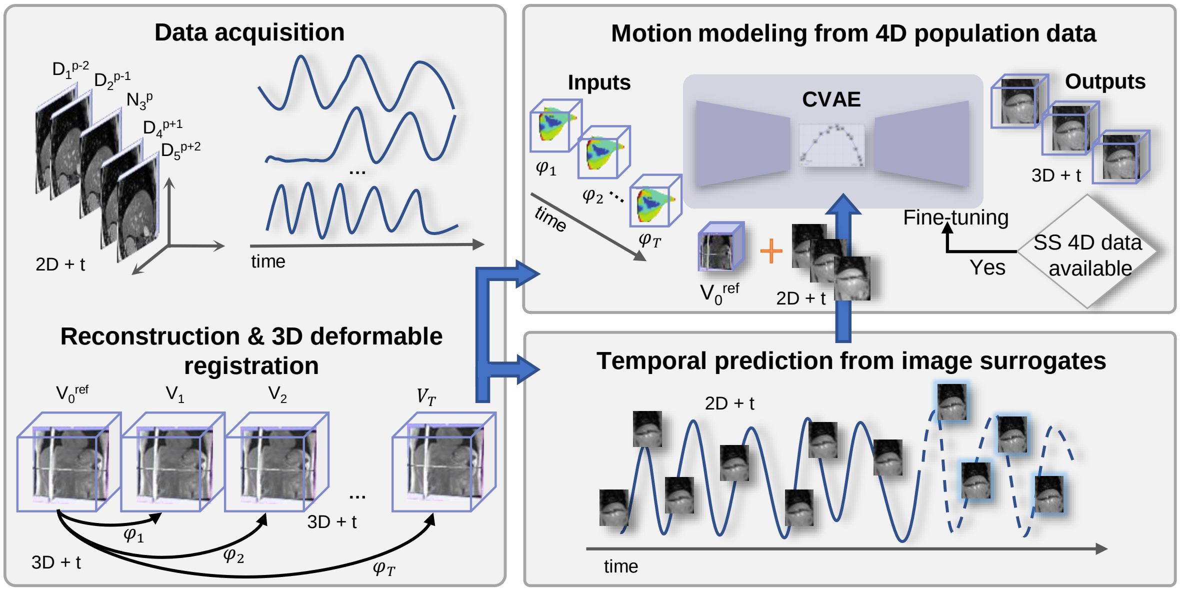


Figure 7.1 Overview of the motion modeling pipeline and its application to IGRT. Cine-MR slices are acquired during free-breathing and then reordered to construct temporal volumes. The organ motion between a reference volume (at a fixed time) and the other temporal volumes is estimated via deformable registration. This data, captured over a population, is employed to fit a probabilistic motion model, which can be readily personalized (via fine-tuning) when a subject-specific (SS) 4D dataset is available before treatment.

a low-dimensional space containing compact representations of respiratory states. Simultaneously, these representations are linked to the surrogate 2D image sequence and subject-specific features extracted from the volumetric reference image. The main contributions can be summarized as follow:

- A unified conditional generative framework, which integrates anatomical information and a history of partial observations as predictive variables for the motion modeling task (Section 7.4.3).
- A temporal predictive mechanism acting on low-dimensional features to forecast multiple future volumes in one shot (Section 7.4.4).
- Demonstration of motion modeling and multi-time prediction capabilities with multiple imaging modalities (MRI, ultrasound) and settings (both population and subject-specific) showing superior performance and advantages over state-of-the-art approaches (Section 7.5).

Considering the rapid evolution of the radiation systems and the impressive performance of deep learning for medical imaging task, this work takes a step towards the introduction of a deep probabilistic motion model that might impact the next generation of image-guided radiotherapy. To facilitate further research and encourage other researchers to build upon our results, the source code of our probabilistic motion model is made publicly available at *https://github.com/lisetvr/4d-dmm*.

## 7.3 Data acquisition and 4D volume reconstruction

Figure 7.1 illustrates our pipeline for the 4D motion model construction and online application for IGRT using 2D surrogates. Acquiring ground truth motion data is the first step for building respiratory motion models, showing the moving organs in free-breathing over time (3D + t) with sufficient temporal resolution (a couple of volumes per second) [219]. In this section we present details on the acquisition protocol and the reconstruction process to create the in-house datasets.

### 7.3.1 MRI datasets

Free-breathing sagittal slices were acquired on 25 volunteers, each providing their written consent. The study was conducted under approval of the CHUM's research ethics committee. The acquisitions were carried out on a 3T Philips Ingenia whole body MRI scanner using a 2D T2-weighted Balanced Turbo Field Echo (bTFE) sequence. This sequence enabled good vessel visualization without using any contrast agent. Image dimensions were $32 \times 176 \times 176$, pixel spacing was $1.7 \times 1.7$ mm$^2$ and the slice thickness was 3.5 mm. An alternating scheme was followed to acquire data frames covering the right liver lobe interleaved with navigator frames taken at a fixed anatomical position, chosen in the middle of the liver. In order to produce time-resolved volumes, the navigator slices were non-rigidly aligned to a pre-selected segmented master navigator. This step was initialized with a rigid transformation and performed using Elastix as the registration framework [214]. The deformation field was parameterized by cubic B-splines at three resolutions. The similarity measure was the normalized mutual information. The liver area inside the master navigator was manually segmented by a specialist, thereby yielding a binary mask. The 2D deformation fields inside the binary mask were used to calculate a similarity cost function to drive the slice stacking as detailed in [2]. By considering only the deformation within the mask, the organ's respiratory motion is isolated from other sources of motion. The core idea behind this reconstruction method is that, whenever the respective preceding and subsequent navigator frames corresponding to two different anatomical slice positions are similar, both data slices will be at the same respiratory state and, therefore, can be stacked together to construct the volume [2]. It should be considered that potential uncertainties during the image acquisition and the subsequent deformable registration may cause artifacts during the sorting process. Nevertheless, the motion observed between temporal volumes still represents a valid ground truth to build the motion model.

The temporal resolution of the volumes was of 450 ms. For each of the 25 cases, 80 different sequences of 2D navigators showing different breathing patterns were acquired with 31 time points each, yielding a dataset of 62 000 volumes. For each volunteer, these 80 sequences showed different motion amplitudes and frequencies as illustrated in Figure 7.2, which portrays the considerable inter-cycle variability that must be taken into account to increase the robustness of the motion model during radiotherapy. Therefore, we leverage this variability as a data augmentation strategy during the dataset creation.

A second free-breathing MRI dataset was acquired from 11 patients diagnosed with hepatocellular carcinoma (6 male, 5 female, ages 70±11). The study was approved by the CHUM's research ethics committee, and patients provided informed consent to participate. The data was acquired continuously during 3 minutes on a Philips Ingenia 3T clinical MRI scanner using a 3D stack-of-stars gradient-echo radial sequence (flip angle=12°, TR=3.4ms, TE=1.4ms, FOV=$450\times450\times250$ mm$^3$, spatial resolution=$1.5\times1.5\times5.0$ mm$^3$) with golden-angle sampling scheme. This sampling scheme used an angular increment of $\approx 111.25°$ between consecutive spokes, which enabled the extraction of a respiratory signal used to reconstruct 10 respiratory phases. For further details on the XD-GRASP reconstruction technique, see [98,99]. For each patient, tumors exceeding 10 mm in the right liver lobe were annotated by an experienced abdominal radiologist using previous diagnostic images. As a pre-processing step for model deployment, volumes were resampled to the spatial resolution of the previous dataset and cropped to $32 \times 64 \times 64$ to focus on the liver and remove organs in the bottom part of the abdomen such as the stomach, pancreas, kidneys and intestines. Henceforth, we will refer to the volunteer and patient datasets as V-MRI and P-MRI, respectively.

### 7.3.2 Ultrasound dataset

Another dataset of free-breathing 4D US sequences from 20 volunteers, who provided their written consent, was acquired using a Philips EPIQ 7G ultrasound system with a X6-1 matrix array transducer. This study was approved by the CHUM's research ethics committee. During acquisition, the ultrasound probe was placed under the sternum along the sagittal plane, capturing a cross section of the left liver lobe. The imaging depth was set to 12cm. Focus and contrast were adjusted to provide the best visualization of the liver and its vessels. Limited to a 15s acquisition window, it was possible to capture up to 3 respiratory cycles with a 250 ms temporal resolution. The acquired volumes were first pre-processed by applying a Bayesian non-local means filter for speckle removal. Then, the volumes were resampled to a $2.0 \times 2.0$ mm$^2$ spatial resolution with a slice thickness of 1.0 mm and cropped to a volume size of $32 \times 64 \times 64$.

## 7.4 Proposed population-based motion model

In the field of respiratory motion modeling, one is often interested in the changes undergone by the organ with respect to a reference breathing phase. Having the 4D ground-truth data, the next step is to fit a model that establishes a relationship between specific surrogates and the motion of interest (see Figure 7.1). Dimensionality reduction and manifold learning theory has shown to be key components for the analysis of organ motion in medical datasets.

Our conditional probabilistic framework learns from population motion data covering a significant variety of breathing patterns. It receives as input: a pre-treatment volume gated at a certain reference respiratory phase and a sequence of image surrogates to estimate the deformation from a previously learned motion distribution. The temporal information of the surrogate sequence acts as a predictive variable to recover the dense displacement vector field (DVF) corresponding to $n$ future respiratory phases. Such conditioning factors are feature vectors forecasted from the latest acquired images.

Figure 7.3 shows a schematic representation of the proposed model which has a different configuration in training and testing phases. The training framework is composed of the following blocks: (1) a DVF inference network for motion estimation, (2) a conditional variational autoencoder to learn the DVF distribution with respect to the surrogate, and (3) multi-time predictive module. During testing, the first component and the motion encoder are removed. Therefore, the motion prediction only depends on sampling the latent distribution and conditioning it on the available intra-treatment partial observations. In the following, we formalize the modeling task, focusing on its application for real-time image-guided motion estimation. We then thoroughly explain the model components and provide its implementation details.

### 7.4.1 Problem formulation

We consider a set of $P$ time-resolved 3D acquisitions spanning $k$ respiratory cycles generating $T$ temporal volumes. For each subject's dataset $p \in P$, a volume $V_{t=r}$ at certain respiratory state $r$ is selected. This reference volume is used to measure relative displacements to the other temporal volumes in a given dataset. The motion observed between a moving image

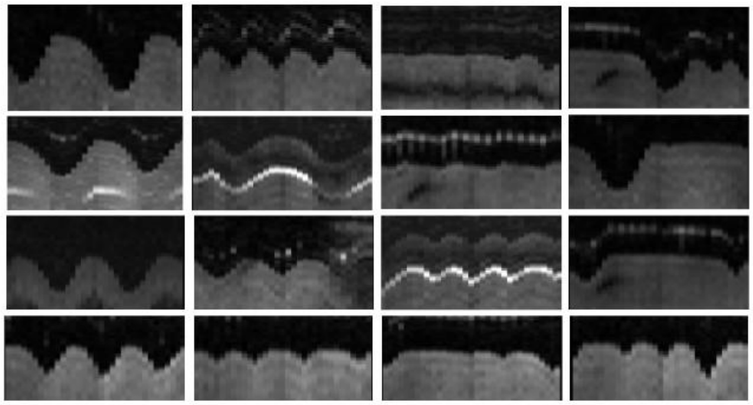


Figure 7.2 Sample breathing patterns captured by the navigator amongst a volunteer population: shallow, deep, regular, irregular. The patterns in the last row are from the same subject.

$V_r \in \mathcal{R}^{H \times W \times D}$ and a fixed image $V_t \in \mathcal{R}^{H \times W \times D}$ is described by a dense displacement field $\phi_t \in \mathcal{R}^{H \times W \times D \times 3}$, where $t \in [1, T] \neq r$, and $H, W$ and $D$ denote the height, width and depth of the volumes, respectively. Hence, each dataset $p$ is comprised by an ensemble of 3D deformations $\Phi^p = \left\langle \phi_1, \phi_2, \ldots, \phi_{T-1} \right\rangle$. The first goal of the motion modeling task consists of mapping each deformation to a low-dimensional space $\phi_t \rightarrow z_t \in \mathcal{R}^d$ where $d \ll H \times W \times D \times 3$, thereby summarizing the observed inputs in a compact representation. Simultaneously, each deformation at time $t$ can be associated with a partial observation, e.g., a 2D slice $I_t \in \mathcal{R}^{H \times W}$. Moreover, to meet the temporal requirements the deformations must be forecasted ahead-of-time. Therefore, the second goal is to relate the partial observations with the correspondent extrapolated-in-time dense deformations.

### 7.4.2 DVF inference network

The motion measurement step using deformable registration constitutes the first step in the motion modeling workflow. Since our method does not rely on any surface-based information (i.e. prior segmentation) and avoids explicit voxel generation, we work with deformations between pairs of volumes from same patients in a dataset. We use a registration function, parameterized with a neural network, which receives a specific reference volume $V_{ref}$ and a target volume $V_t$ at time $t$ as inputs to generate the breathing-induced organ DVF matrix $\phi_t$ between them. In our experiments, $V_{ref}$ is taken at the end-exhale phase since it presents the most reproducible liver shape [241]. Parameterizing the deformation through a neural network enables a single differentiable end-to-end pipeline. Furthermore, it requires less memory and the inference is faster than traditional registration methods. These represent

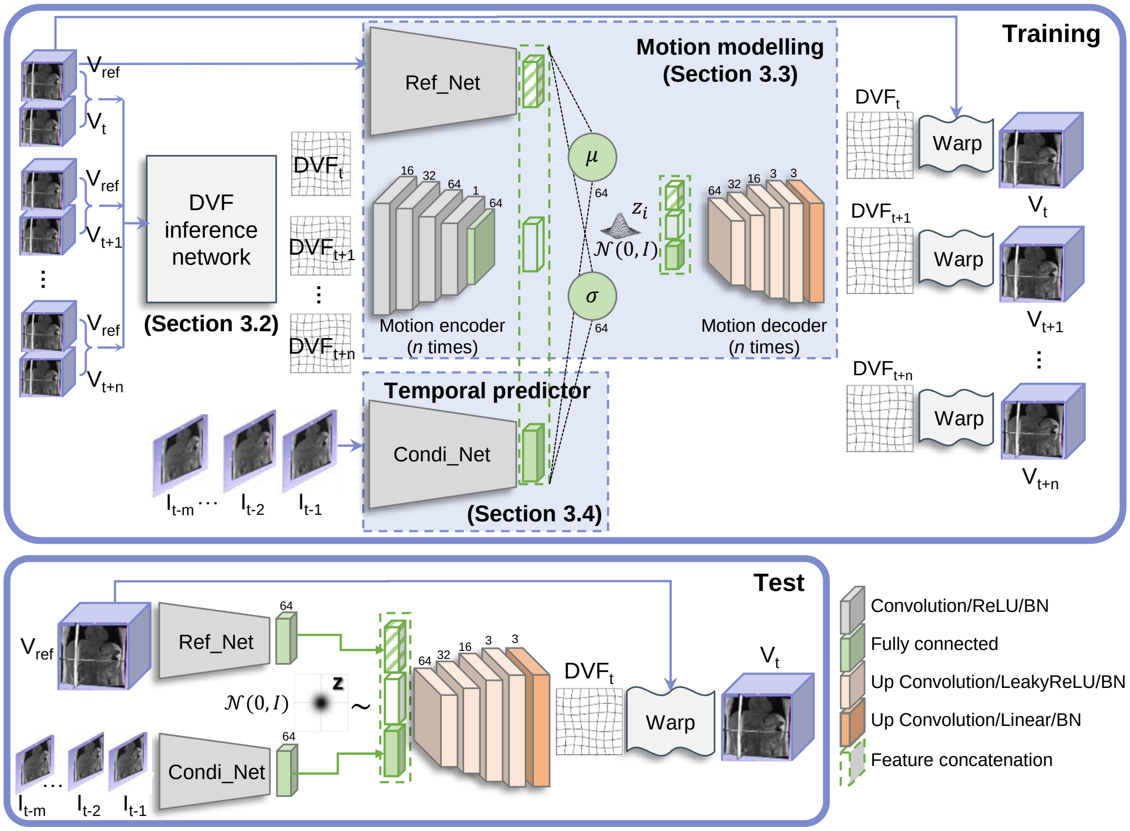


Figure 7.3 Schematic representation of the proposed probabilistic motion model. Top: During training, the inputs are: a reference volume ($V_{ref}$) and a set of target volumes $\{V_t, V_{t+1}, \cdots, V_{t+n}\}$ at $n$ time steps. The deformation between each pair of volumes, i.e. $V_{ref}$ and $V_i$, are estimated through a pre-trained inference DVF network. These deformations and the inputs volumes are fed to a multi-branch convolutional neural network composed of three branches: (1) an auxiliary encoder that receives the reference volume, namely Ref-Net; (2) a motion encoder, which is repeated according to the amount of input deformations ($n$ times), and (3) a temporal predictive network, which outputs the extrapolated-in-time feature vectors used as conditioning variables, namely Condi-Net. The outputs of each branch are combined together according to each time. Then it is constrained to form a Gaussian distribution, conditioned on the predictive variables. The decoder generates a DVF from each input feature vector, meaning a phase-specific dense 3D deformation. Bottom: During testing, given the partial observations and the 3D anatomical reference, it is possible to sample the latent space and recover the 3D deformations.

important advantages during training.

We assume that both volumes were previously rigidly aligned to a common reference space, meaning that their origin and orientation coincide. This step could be performed with traditional registration frameworks such as Elastix [214] or Plastimatch (*www.plastimatch.org*). For the deformable registration, we employ the U-net-like architecture proposed by [42] with pre-trained weights, which is a well-validated structure. This means that the DVF inference network was previously trained, thus during model optimization their weights remain static. It should be noted that the proposed motion modeling framework is agnostic to the approach used for deformable registration.

### 7.4.3 Conditional motion modeling

In our case, the set of temporal volumes, acquired under free-breathing, differs from one another by the tissue deformation due to complex respiratory motion. These deformations lie in a high-dimensional space, determined by the number of voxel and motion components. However, the spatio-temporal variation is caused by a much smaller number of degrees of freedom, and hence the underlying structure can be captured by a low-dimensional subspace. Therefore, it can be seen as a set of points on a manifold of many fewer dimensions. Such embeddings can be uncovered using the capacity of autoencoders to learn a non-linear parametric mapping from volume deformations to their latent representations.

Let $\Phi = \{\phi_{t+1}, \phi_{t+2}, \ldots, \phi_{t+n}\}$, $I_{seq} = \{I_t, I_{t-1}, \ldots, I_{t-m}\}$ and $V_{ref}$ be the sequence of $n$ future 3D deformations, the sequence of 2D slices at times $t$, $t+1$ and $t-m$, and the reference volume, respectively. The goal is to maximize the conditional probability distribution $p_{true}\left(\Phi|I_{seq}, V_{ref}\right)$ of obtaining the sequence of deformations $\Phi \in \mathcal{R}^{H\times W\times D\times 3\times n}$ given the available partial information and subject anatomy, where $H, W$, $D$ and $n$ are the height, width, depth of the volume and number of predicted time steps, respectively. Accordingly, we aim at learning a parameterized model with parameters $\theta$ to sample new phase-specific deformations, similar to samples from the unknown distribution $p_{true}$. This dependency, i.e. $p_\theta\left(\Phi|I_{seq}, V_{ref}\right)$, can be expressed by the law of total probability, which relates the conditional and marginal probabilities:

$$p_\theta\left(\Phi|I_{seq}, V_{ref}\right) = \int_z p_\theta\left(\Phi|z, I_{seq}, V_{ref}\right) p(z)\, dz \tag{7.1}$$

where the likelihood $p_\theta\left(\Phi|z, I_{seq}, V_{ref}\right)$ is chosen to be a Gaussian distribution, which is continuous in $\theta$. However, solving this integral over $z$ would require a very large number of samplings, which is not viable. Therefore, to compute the left term of Eq. 7.1, only $z$ values likely to produce $p_\theta\left(\Phi|I_{seq}, V_{ref}\right)$ are considered, namely the posterior $p_\theta(z|\Phi, I_{seq}, V_{ref})$.

Because computing the exact posterior distribution of the CVAE is analytically intractable [242], it is approximated through a distribution $q_\psi\left(\cdot\right)$ with parameters $\psi$. By using Bayes' theorem, we have:

$$\begin{aligned} E_{z\sim q}[\log p_\theta(\phi_i|z_i, I_{seq}, V_{ref})] = E_{z\sim q}[\log p(z_i|\phi_i, I_{seq}, V_{ref})] \\ - \log p(z_i|I_{seq}, V_{ref}) + \log p_\theta(\phi_i|I_{seq}, V_{ref}). \end{aligned} \tag{7.2}$$

After subtracting $E_{z\sim q}[logq_\theta(z)]$ from both sides and rearranging the terms, the approximated posterior can be related to the true posterior through the Kullback-Leibler (KL) distance:

$$\begin{aligned} \log p_\theta(\phi_i|I_{seq}, V_{ref}) - KL[q(z_i|\phi_i, I_{seq}, V_{ref})||p(z_i|\phi_i, I_{seq}, V_{ref})] = \\ E_{z\sim q}[\log p_\theta(\phi_i|z_i, I_{seq}, V_{ref})] - KL[q(z_i|\phi_i, I_{seq}, V_{ref})||p(z_i)]. \end{aligned} \tag{7.3}$$

With a high-capacity function to approximate the posterior, the second term of the left side is expected to be negligible in the ideal case. Therefore, maximizing $\log p_\theta(\phi_i|I_{seq}, V_{ref})$ is equivalent to maximizing the evidence lower bound (ELBO) on the right side of Eq. 7.3, which basically contains the expectation of the reconstruction term and the KL distance between the prior and the approximated posterior. Generally, the prior is assumed to be a multivariate Gaussian distribution with covariance $I$, i.e. $p_\theta(z_t) \sim \mathcal{N}(0, I)$, since it can be computed in a closed form and is differentiable.

In practice, an encoder network is adopted to find the approximation:

$$q_\psi(z_i|I_{seq}, V_{ref}) \; = \mathcal{N}\left(\mu(\phi_i, I_{seq}, V_{ref}), \, \sigma(\phi_t, I_{seq}, V_{ref})\right). \tag{7.4}$$

This network, parameterized with stacked 3D convolutional layers, learns the mean $\mu \in \mathcal{R}^d$ and diagonal covariance $\sigma \in \mathcal{R}^d$ from the data, as illustrated in the upper middle of Fig. 7.3. At training, the sampling of $z_i$ is differentiable with respect to $\mu$ and $\sigma$ by using the *reparameterization trick* [242], and defining $z_t = \mu + \epsilon * \sigma$, where $\epsilon \sim \mathcal{N}(0, I)$.

Following the methodology of the CVAE, the distance between both distributions $p_\theta$ and $q_\psi$ is minimized using the KL-divergence. This loss term is inserted within the total loss function, which also aims at minimizing a reconstruction loss. Unlike autoencoders, we avoid calculating voxel-wise differences over motion fields. Instead, we implicitly regress $\phi_i$ through image similarity. Hence, in the spatial warping block, the reference volume is warped (denoted with the symbol $\circ$) with the transformation provided by the decoder enabling the model to calculate a reconstruction term $\mathcal{L}_{rec}$ between $V_{ref} \circ \phi_i$ and the expected in-room volume $V_i$. We use stochastic gradient descent to find the optimal parameters $\hat{\theta}$ by minimizing the following loss function:

$$\arg\min_{\theta}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{rec}\left(V_{ref}\circ\hat{\phi}_t, V_t\right)+\mathrm{KL}\left(q_{\psi}(z_i|I_{seq},V_{ref})||p_{\theta}(z_i)\right)\right] \tag{7.5}$$

where the KL-divergence can be computed in closed form.

The reconstruction term, $\mathcal{L}_{rec}$, computes the similarity between the estimated voxel output against the target voxels $(V_i)$ while ensuring smooth motion fields using a diffusion regularizer on the spatial gradients:

$$\mathcal{L}_{rec}=\mathcal{L}_{sim}\left(V_i,\hat{V}_i\right)+\alpha\mathcal{L}_{smooth}\left(\hat{\phi}_i\right) \tag{7.6}$$

where $\hat{V}_i$ results from warping $V_{ref}$ with the estimated motion $\hat{\phi}_i$, $\alpha$ is a regularization parameter, and $\mathcal{L}_{smooth}\left(\hat{\phi}_i\right)=\sum_{p\in\mathbb{R}^3}\|\nabla\phi\left(p\right)\|^2$ computes the differences between neighboring 3D positions $p$.

### 7.4.4 Multi-time predictive network (Condi-Net)

For the design of real-time motion compensation systems, the involved latencies must be considered. Therefore, the goal of the multi-time predictive network is to enable future deformation recovering. This task is conducted by the conditional branch ("Condi-Net" block

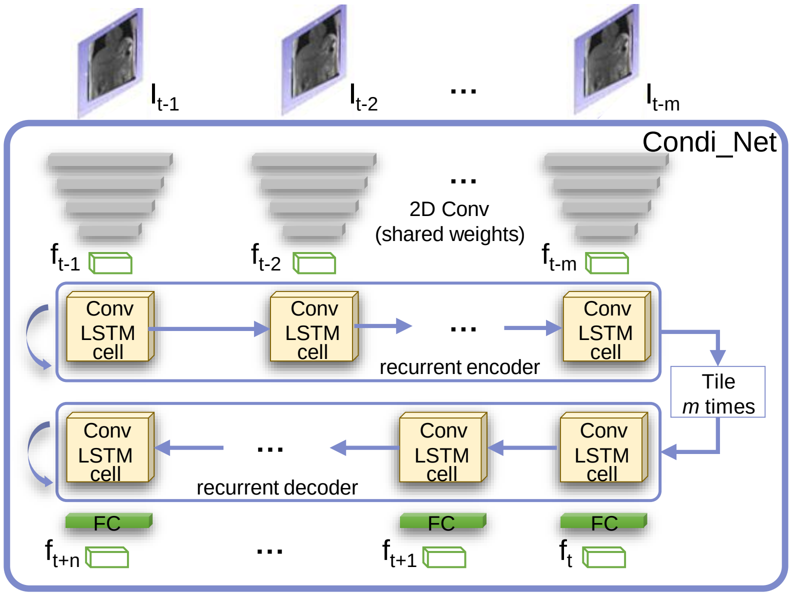


Figure 7.4 Schematic representation of the temporal predictive network, which receives an input image sequence and outputs the extrapolated-in-time feature vectors used as conditioning variables.

in Figure 7.3). This sub-network tries to maximizes the conditional probability of predicting a sequence of feature vectors $h_i$, corresponding to future time steps, given the input images, i.e., $P(h_t, h_{t+1}, \cdots, h_{t+n}|I_{seq})$.

Figure 7.4 depicts the internal configuration of the "Condi-Net" block. It receives a surrogate 2D image sequence ($I_{seq}$), where each temporal image is concatenated along the feature dimension with their corresponding slice in the reference volume. Ideally, this additional information is a common reference that helps with the respiratory phase discrimination. Subsequently, each bi-channel image is passed through a feature extraction function, parameterized with convolutional neural networks. The resulting feature representations are then concatenated to form a new temporal dimension. The temporal feature representations is then forecasted through recurrent cells arranged in an encoder-decoder configuration. Such a design is inspired by the *seq2seq* mechanism, widely used for natural language processing (NLP) and other related time-series tasks [148]. The encoder receives the spatio-temporal features and outputs a single vector, known as context vector. The hidden state from the final encoder cell is an embedding containing a condensed representation of the sequence. It is tiled and fed to the decoder, which learns how to extrapolate the feature vectors associated to future time steps.

### 7.4.5 Implementation details

In the proposed architecture, we develop a multi-branch convolutional neural network composed of three branches: (1) the motion encoder, (2) the auxiliary encoder for the reference volume ("Ref-Net" sub-network) and (3) the image sequence encoder, which enable temporal predictions ("Condi-Net" sub-network) (see Figure 7.3). The first and second branch possess identical configurations except by the number of input channels. In the first case, it receives the 3 channels pertaining to the motion components. In the second case, it receives a single channel with the voxel intensity of the reference volume. Both encoders are composed of successive pairs of 3D convolutions with a kernel size $3 \times 3 \times 3$. The inclusion of the subject-specific anatomical 3D information seeks to alleviate the ill-posedness of the problem. The first layer in the stack has stride of 2 and the second has single stride. In all cases they are followed by ReLU activations and batch normalization (BN). The motion encoder is repeated $n$ times, i.e., the same number of output volumes. Each one of them shares their weights.

The third branch receives a sequence of bi-channel images. Each of them is passed through a shared stack of 2D convolutions with kernel size $3 \times 3$ and a stride of 2, followed by ReLU activations and batch normalization. Subsequently, the feature representations are concatenated forming a new temporal dimension, which is fed to the recurrent encoder-decoder configuration. In the model, this configuration is implemented with convolutional long short-term memory (ConvLSTM) with 64 feature maps and kernel size 3 × 3. The encoder consists of $m$ cells, depending on the number of temporal slices available. Similarly, the decoder consists of $n$ cells, depending on the desired number of output volumes.

Each of the aforementioned model branches, i.e., the motion encoder, "Ref-Net" and "Condi-Net", ends in a fully connected (FC) layer with equal size to the latent dimension. Their respective outputs are further concatenated and mapped to two additional FC layers to generate $\mu$ and $\sigma$, which are combined with $\epsilon$ to construct the latent space sample $z_i$, representing the normal Gaussian distribution. The rationale behind this concatenation scheme is to create the distribution over the latent vector considering the patient anatomy and spatio-temporal consistency associated to the 3D deformation.

Once the latent distribution is created, the conditional dependency is explicitly modeled by the concatenation of $z_i$ with the feature representation of $V_{ref}$ and each one of the extrapolated embeddings $h_i$. The resulting vector is then fed to the decoder, which learns how to map it back into the input space, thereby yielding a reconstructed displacement field $\hat{\phi}_t = g_\theta(z)$. The decoder is modeled with a stack of convolutional neural networks, with kernel size 3×3×3 and interleaved strides of 1 and 2 to upsample the spatial dimensions while gradually decreasing the number of channels. These convolutional layers are followed by Leaky ReLU activations (0.2) and batch normalization, except the last one, which has a linear activation. It is important to note that the decoders used to generate the different temporal deformations share their weights. While one may consider leveraging 4D convolutions to manage the volumetric sequence, this would change the approach since the decoder can be seen as an unique generator that maps from the low-dimensional motion model to the high-dimensional space. Finally, the reference volume is resampled according to each generated temporal deformation using a differentiable spatial warping layer (STN) [161]. Using this scheme, our model is able to provide volumetric information.

At test time, the encoder is disabled since it is assumed that there will be no volumetric motion information. Hence, the decoder operates as a generative network given only the patient anatomy and the in-room cine acquisition, yielding realistic DVFs by sampling $z_i \sim \mathcal{N}(0, I)$.

### 7.4.6 Training and personalization protocol

The network's parameters were optimized using the Adam optimizer with an initial learning rate ($lr$) set at $10^{-3}$, which was reduced by a factor of 2 after each sequence of 3 epochs without improvement. For model personalization, the $lr = 10^{-5}$ was progressively reduced after 3 epochs without improvements in the validation loss measured on the subject-specific volumes. Mean centering and standard deviation normalization were applied to the input images and volumes. This normalization was performed on a per-volume basis. Training was performed in PyTorch with a batch size of 10 to exploit GPU acceleration. We used a leave-one-out validation scheme, considering a different anatomical case for testing. We adopted a negative local cross correlation (NCC) as similarity loss function when working with the MRI dataset. On the other hand, for the ultrasound (US) dataset, we used a similarity loss based on Mean Squared Error (MSE), which was helpful for the model convergence. In equation 7.6, $\alpha = 1.0$ and $\alpha = 0.01$ when $\mathcal{L}_{sim}$ was NCC and MSE, respectively.

## 7.5 Experiments and results

In this section we present a series of experiments in order to analyze the impact of each component in the form of ablation study, justify the choices made in the model design and understand the internal structure of the latent space for the motion modeling task (Section 7.5.2). Moreover, in Section 7.5.3 we quantitatively assess the performance of the proposed framework under two possible scenarios: (1) when a prior subject-specific 4D data is not available and thus a population model is applied, and (2) when a subject-specific (SS) 4D acquisition is available before treatment and thereby used to personalize the model.

### 7.5.1 Validation methodology

In the first scenario (i.e. unavailable subject-specific data), the predictive capability of the proposed model, tested on unseen cases, was compared to several approaches that have been introduced in the context of IGRT: (1) a motion extrapolation (ME) based method [202], which is based on deformable registration between interleaved 2D images and their corresponding slices in the pre-treatment 3D volume; (2) a deep neural network that combines feature representations from a reference volume and a surrogate slice to generate a 3D deformation [213] (denoted as DN, which stands for deep network); (3) a model based on motion autoencoding (AE), which aims at minimizing the distance between surrogate images and 3D deformations in a low-dimensional space [190]. In addition, we evaluate the effect of the anatomical plane on the predictive accuracy. Sagittal and coronal planes were considered since they both capture the cranial-caudal direction, the direction in which the largest respiratory motion is present. The motion model created with MRI data from 24 volunteers (V-MRI dataset) was evaluated both with an unseen anatomy acquired using the same protocol and on a separate dataset of 11 liver cancers patients (P-MRI dataset). Therefore, this independent hold-out test dataset was not used to develop the model or tune the hyper-parameters.

For the second scenario (i.e. available subject-specific data), we compared our model to a subject-specific statistical model coupled with AdaBoost (AB) multilayer perceptron to enable spatio-temporal prediction as detailed in [174]. In this case, the experiments were conducted only with the V-MRI dataset. The 20 minutes were divided for personalization and testing purposes. Our model was first fine-tuned with 620 volumes, corresponding to the first 5 min of the V-MRI dataset, and then tested with the remaining volumes.

We compare the performances based on target tracking accuracy, by measuring the distance between five expert-selected blood vessel and bifurcations in the ground-truth and predicted images. The prediction error for each landmark $l$ in a temporal image $t$ was measured as the Euclidean distance $E_{l,t+\Delta} = \|g_{l,t+\Delta} - p_{l,t+\Delta}\|$ between target $g$ and predicted $p$ positions. We also compute the global registration error using 3D deformable registration between ground-truth and predicted volumes. This was conducted using the multiresolution B-spline transformation model implemented in Elastix [214]. The spatio-temporal prediction error was defined as the voxel-wise vector magnitude of the estimated DVFs. Moreover, we report the geometrical error for all motion states, which is defined as the Euclidean norm of the voxel-wise vector difference between the ground truth and the predicted DVFs. Although the ground-truth DVFs do not necessarily represent the real motion due to errors introduced by the 4D reconstruction process and by the registration algorithm, they still represent a valid ground truth for the motion model. For certain experiments we used the MSE and NCC similarity metrics, as well as Structural Similarity (SSIM), to capture the spatial consistency between ground truth and predicted volumes.

During inference, the reference volume ($V_{ref}$) was excluded from the processed volumes. Hence, it was not considered in the error computation. Furthermore, the reference volume was taken at the very first end-exhale while the evaluated breathing cycles were the last ones. This means that the elapsed time between both was maximized within the limits of the dataset (15-20 min interval for the V-MRI). Statistical significance was calculated by applying a Wilcoxon signed-rank test to reject the null hypothesis. In all the tests, $p < 0.01$ was considered to indicate a statistically significant difference. Effect size was measured using Pearson correlation coefficient ($\rho$).

In Section 7.5.3 we report results on temporal experiments, which aim at: (1) determining which is the best predictive mechanism to get multiple output volumes, (2) documenting model behavior with varying number of input images and output volumes, and (3) reporting the prediction accuracy for a predictive horizon of more than one time step. These experiments were conducted on the V-MRI dataset. Since its temporal resolution is 450 ms, the

Table 7.1 Performance metrics obtained for different variants of the proposed model on the V-MRI dataset. Values are mean ± std. ($95^{th}$ percentile). The p-value (effect size) is reported for the geometrical error distributions between consecutive experiments.

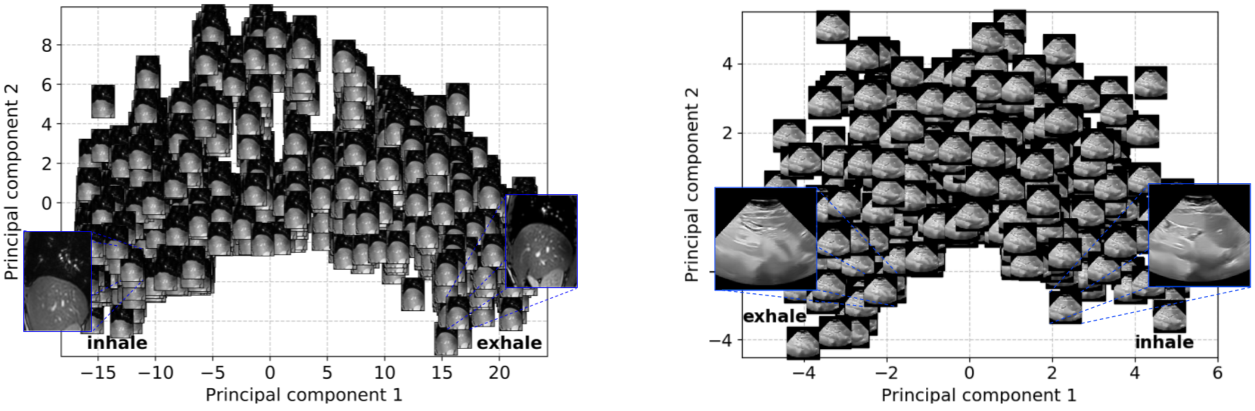
| **Ablation models** | **MSE** | **NCC** | **SSIM** | $\boldsymbol{p}$ **value** |
|---|---|---|---|---|
| CVAE+$I$ | 0.15 ± 0.09 (0.35) | 0.72 ± 0.12 (0.89) | 0.69 ± 0.14 (0.89) | - |
| CVAE+$(I\|I_{ref})$ | 0.15 ± 0.10 (0.35) | 0.74 ± 0.13 (0.91) | 0.71 ± 0.14 (0.90) | $\ll 0.01$ ($\rho$=0.83) |
| CVAE+$(I\|I_{ref}) + V_{ref}$ | 0.14 ± 0.08 (0.32) | 0.74 ± 0.12 (0.91) | 0.72 ± 0.13 (0.91) | $\ll 0.01$ ($\rho$=0.89) |
| **CVAE+$\boldsymbol{(I_s\|I_{ref})+V_{ref}}$** | **0.12 ± 0.07 (0.27)** | **0.77 ± 0.11 (0.93)** | **0.75 ± 0.12 (0.93)** | $\ll 0.01$ ($\rho$=0.81) |

times associated with the predictive horizons $n = 1, 2, 3, 4, 5$ are $450, 900, 1350, 1800, 2250$ ms, respectively. Finally, in Section 7.5.4 we present qualitative results on the employed datasets.

### 7.5.2 Ablation study

Table 7.1 presents the results from several metric for the different model variants, which were trained under the same conditions to predict the next temporal volume. These values were computed on the V-MRI dataset. The ablation study starts with a baseline architecture, which is composed solely of the conditional variational autoencoder and a single 2D image. In the second variant, the surrogate image was concatenated with the correspondent slice in the reference volume ($I_{ref}$). The appearance of this slice remains invariant across all the images belonging to a same subject. Therefore, it is useful to identify and discriminate the respiratory phases. In the third case, the anatomical information was incorporated by encoding the reference volume ($CVAE$+$(I|I_{ref})$+$V_{ref}$) and exploiting their feature during the volume generation. In the last version, we leveraged the spatio-temporal information provided by an image sequence ($I_s$) where each temporal image is concatenated with the reference slice. It can be observed that all the similarity metrics reflects a gradual improvement between consecutive versions. In all cases, the differences between consecutive experiments were found to be statistically significant, with $p \ll 0.01$. The large effect sizes reported in Table 7.1 correspond to the SSIM metric.

To investigate the structure of the latent space, we applied principal component analysis on the latent code vectors to reduce their dimensionality to a single point in a bidimensional Cartesian space. The manifolds shown in Figure 7.5 for MRI and US motion models reveal that data points span through the respiratory cycle. Besides, the points corresponding to



(a)

(b)

Figure 7.5 Low-dimensional mapping visualization (in 2D) of the latent representation in (a) V-MRI and (b) US datasets.

volumes at end-exhale and end-inhale were found at separate ends of the manifold. This phase discrimination explains the learning process and is plausible for a motion model.

### 7.5.3 Quantitative results

#### Single time-point experiments

**Inter-cycle tracking errors** Given the important inter-cycle variability and irregular breathing patterns captured in the volunteer MRI dataset, as shown in Figure 7.2, a single blood vessel was tracked through several cycles to analyze how the model copes with these effects. For the sake of clarity and visualization, a single trajectory will be shown. This blood vessel's location was selected at the medial position and near to the diaphragm, since this is the area with the largest amplitude of movement.

Figure 7.6 displays the target and predicted relative vessel displacements (in mm) in the superior-inferior (SI) and anterior-posterior (AP) directions in three cases with irregular breathing pertaining to the volunteer MRI dataset. The graphs also present the associated errors for all subjects, which are lower than 2 mm and 1 mm for SI and AP motion planes, respectively. The proposed model demonstrates an acceptable consistency with the target trajectory even in the presence of involuntary breath-holds and variable cycle amplitudes. The ability of coping with these cycle-to-cycle variations is essential to ensure predictive robustness when deployed in the clinical setting.

**Target tracking** Table 7.2 reports the geometrical accuracy (for the next time step) between ground-truth and predicted landmark positions for the proposed model working both in

Table 7.2 Target tracking errors (in mm) measured at selected respiratory phases for the V-MRI dataset. These values were measured for the next time step, i.e. a horizon of 450 ms. Overall values consider all the phases. Values are mean ± std ($95^{th}$ percentile).

| **Model** | **Mid-inhale** | **End-inhale** | **Mid-exhale** | **End-exhale** | **Overall** |
|---|---|---|---|---|---|
| Initial m. | 9.5±4.5(15.6) | 11.6±5.6(18.9) | 4.4±4.0(12.1) | 1.6±3.6 (8.5) | 6.7±4.4(13.7) |
| DN [213] | 4.8±3.1(12.4) | 5.3±4.5(10.3) | 3.1±1.5 (5.5) | 2.6±2.0 (4.7) | 3.9±2.7 (8.2) |
| ME [202] | 3.0±2.7 (8.8) | 2.5±2.5 (6.1) | 2.5±1.7 (4.6) | 1.9±1.8 (4.3) | 2.4±2.0 (5.8) |
| AE [190] | 2.9±2.5 (6.0) | 3.3±2.9 (5.7) | 2.3±1.6 (4.4) | 2.1±1.7 (3.1) | 2.6±2.1 (4.8) |
| Proposed (sag, P) | 2.9±2.3 (5.9) | 3.1±2.5 (5.0) | 2.4±2.3 (4.9) | 2.3±1.9 (3.3) | 2.6±2.2 (4.7) |
| **Proposed (cor, P)** | 2.4±2.0 (3.8) | 2.9±2.2 (4.5) | 2.1±1.5 (3.5) | 2.0±1.9 (2.7) | 2.3±1.9 (3.6) |
| PCA+AB [174] | 1.6±2.0 (3.6) | 2.0±2.6 (4.7) | 1.6±0.9 (2.9) | 2.0±1.2 (3.2) | 1.8±1.6 (3.6) |
| **Proposed (cor, SS)** | **1.4±1.1 (3.1)** | **1.8±1.6 (4.1)** | **1.3±1.0 (3.1)** | **1.1±0.8 (3.0)** | **1.4±1.1 (3.3)** |

population and subject-specific modes, as well as for related approaches. Furthermore, these tracking errors are reported at different phases through the respiratory cycle. As a reference,

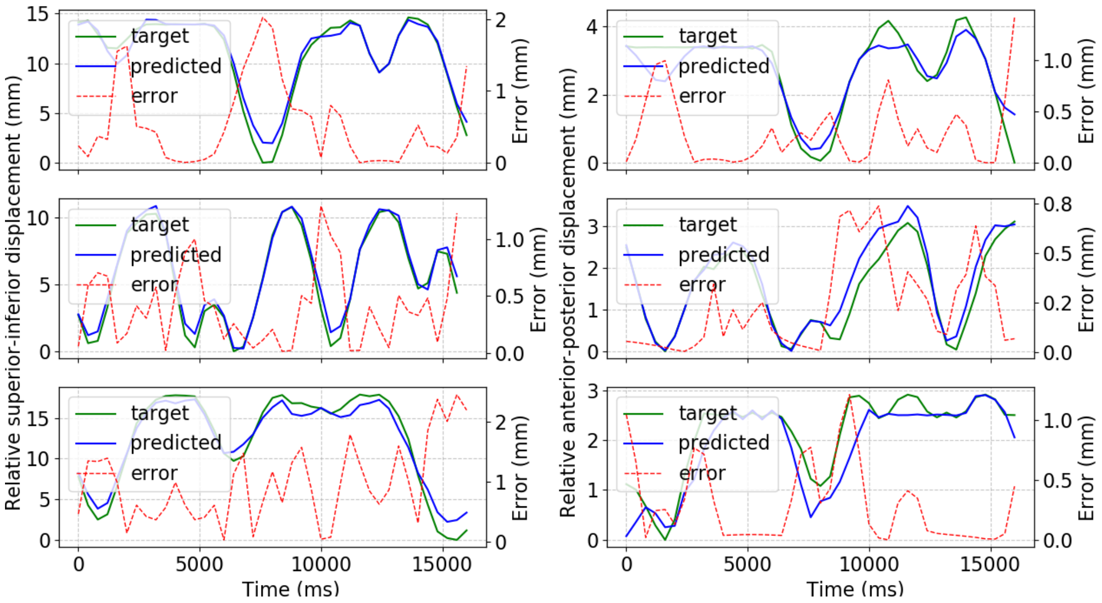
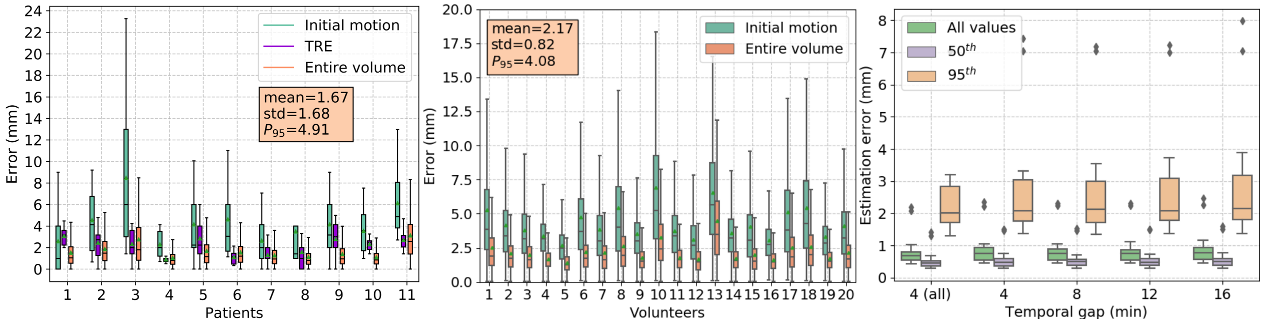


Figure 7.6 Vessel trajectories in the superior-inferior and anterior-posterior motion planes observed in three subjects with irregular breathing in the V-MRI dataset. Dashed red lines represent the error (in mm).



(a) (b) (c)

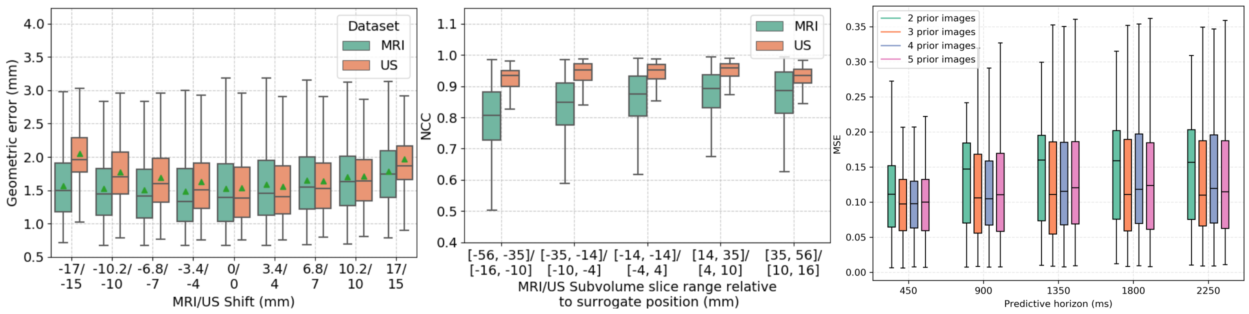
Figure 7.7 Estimation errors (in mm) considering the whole volume for (a) P-MRI and (b) US datasets. (c) Analysis of the drift effect on the estimation error when increasing the temporal gap between training and test subsets in the V-MRI dataset.

the first row contains the errors measured when there is no motion compensation (Initial motion). The values reveal that using coronal plane slices yield an increased performance compared to the sagittal view. This can be attributed to the fact that the coronal plane covers a larger area of the organ than the sagittal plane. Moreover, we found the differences between measurements obtained using sagittal and coronal slices as being statistically significant ($p \ll 0.01$, $\rho = 0.89$). It can be observed that the most challenging predictions were near the end-inhale phase, which is well-known to be prone to inter-cycle variability.

The proposed model, driven by coronal slices, demonstrates the ability to predict deformations throughout all the respiratory cycle with a mean overall error of 2.3 mm in unseen cases. This represents an improvement of 1.6 mm, 0.1 mm and 0.3 mm to DN, ME and AE approaches, respectively. Additionally, when applied to the subject-specific configuration, the overall error decreased to 1.4 mm, a sightly better performance than the PCA+AdaBoost approach. Therefore, our model achieves state-of-the-art results for target tracking while introducing new advantages.

Figures 7.7a and 7.7b display the error distribution per case considering all spatio-temporal voxel-wise displacements for the (hold-out) MRI and US datasets. Additionally, Figure 7.7a contains the target registration error for the center-of-mass of each liver tumor or the average if more than one. The values of breathing magnitudes are in the same order as the values reported by other studies on the quantification of liver motion [119]. It can also be observed that the magnitudes exhibit large differences between them, which showcases the inter-subject variability. The overall computed mean error was 1.67 ± 1.68 mm and 2.17 ± 0.82 mm for patient MRI and US datasets, respectively.

**Drift analysis** The drift that undergoes the liver due to respiration has been described



(a) (b) (c)

Figure 7.8 (a) Geometrical error (in mm) when varying the position of the surrogate slice both in V-MRI and US datasets. (b) NCC between GT and predicted volumes at 5 different anatomical positions along the right-left axis both in V-MRI and US datasets. (c) MSE between GT and predicted volumes in the V-MRI dataset when varying the number of prior images in the conditioning sequence.

in several studies [13, 206], which may cause a negative effect on the system accuracy, for instance, in gated treatment with external respiratory signals [243]. Generally, this outcome becomes visible in longer acquisitions. Hence, we used our largest dataset (volunteer MRI) to validate the impact of the organ drift on the subject-specific model. Each dataset (with 20 min duration) was divided into 5 subsets of 4 minutes. The last subset was spared as a common testing set for all the experiments, namely, when the temporal gap and the training data were: (a) 4 min and all the training subsets, (b) 4 min, (c) 8 min, (d) 12 min and (e) 16 min using only the fourth, third, second and first subsets, respectively. It can be seen in Figure 7.7c that the error distributions show a slightly degradation as the temporal gap between training and testing sets increases.

**Surrogate slice positioning** Figure 7.8a presents the model behavior in terms of the geometrical error, when the surrogate slice position is shifted from the original one used during training, while the reference slice was fixed. The reference slice is extracted from the reference volume at the same anatomical position as the surrogate image. Positive values in the $x$ axis correspond to shifts from the middle to one extreme of the volume while negative values correspond to shifts from the middle to the opposite extreme. It can be observed that, for both V-MRI and US datasets, even though the error is slightly increased, the model is still tolerant to the shift. Interestingly, if the reference slice is also shifted to the same position as the surrogate, there is no performance degradation (see graphs in the Supplementary materials). This finding confirms that, as long as the conditional branch is able to identify the phase, the prediction will be satisfactory. This characteristic confers robustness to the model, and represents an important advantage over current techniques.

**Volume's quality** We also compared the spatial consistency between GT and predicted volumes at 5 different sub-volumes along the right-left axis. This experiment focused on the model's robustness towards abrupt changes with regards to predictive quality depending on the volume area. Results reported in Figure 7.8b show a stable similarity across all the US sub-volumes. In the MRI data, the quality is slightly degraded on the leftmost volumes, which generally does not contains the organ of interest.

**Deformation analysis** Finally, the plausibility of the deformations was assessed using the Jacobian matrix determinant ($|J|$). Values of |J| below or equal to zero indicate folding areas resulting from the crossing of the motion vectors. It should be noticed that the analysis is aimed at assessing the quality of the deformations predicted by the model instead of quantifying the accuracy of the deformable registration step. The proposed model obtained a percentage of voxels with a non-negative $|J|$ of $99.3 \pm 1.3\%$ [88.9, 100], $99.9 \pm 0.1\%$ [99.1, 100] and $98.3 \pm 3.4\%$ [83.5, 100] in V-MRI, P-MRI and US datasets, respectively. This suggests that, on average, it yields smooth and invertible deformations. Furthermore, we report the deviations of |J| from unity within the liver. The mean ± std and $95^{th}$ percentile ($P_{95}$) of all the deviations is $0.07 \pm 0.23$ $P_{95} = 0.44$, $0.02 \pm 0.11$ $P_{95} = 0.19$, and $0.02 \pm 0.47$ $P_{95} = 0.68$ for V-MRI, P-MRI and US datasets, respectively. These results show that |J| values are close to unity within the liver, which means anatomically plausible motion fields. Violin plots with the dispersion from unity as well as visual results with the spatial distribution of |J| are included in the Supplementary material.

### Temporal experiments

Table 7.3 shows a comparison based on the geometrical errors between different predictive mechanisms, namely, ConvGRU, 3D convolution and ConvLSTM to process the image sequence and forecast the future deformations, on a time horizon spanning 1.3 s. Each column shows the statistical values that summarize the error distributions obtained for a horizon of

Table 7.3 Geometrical errors (in mm) obtained from the V-MRI dataset with different alternatives of processing the conditional image sequence to extrapolate future times. Values are mean ± std ($95^{th}$percentile).

| Predictor | $\Delta t = 450$ ms | $\Delta t = 900$ ms | $\Delta t = 1350$ ms |
|---|---|---|---|
| ConvGRU | 1.6 ± 0.9 (3.7) | 1.7 ± 1.1 (3.9) | 1.3 ± 1.0 (3.2) |
| 3D Conv | 1.4 ± 1.0 (3.2) | 1.6 ± 1.2 (4.2) | 1.3 ± 0.9 (3.3) |
| **ConvLSTM** | **1.2 ± 0.6 (2.6)** | **1.4 ± 0.9 (3.3)** | **1.3 ± 0.9 (3.1)** |

$n = 3$ time steps. Overall, the model using ConvLSTM yields the lowest errors compared to the other two variants, a result that was found to be statistically significant, $p \ll 0.01$, $\rho = 0.5$ (ConvLSTM/3DConv), $\rho = 0.6$ (ConvLSTM/ConvGRU).

Figure 7.8c presents MSE values between ground-truth and predicted volumes for different predictive horizon $n = \{1, 2, 3, 4, 5\}$ when varying the length of the surrogate image sequence $m = \{2, 3, 4, 5\}$ given to the proposed model. The case with one input image was not considered as it does not provide spatio-temporal information. These results were obtained following a leave-one-out scheme on a subject level using the V-MRI dataset. Generally speaking, the performance increases as more images are provided as input, which is particularly evident for $m = \{2, 3\}$. On the other hand, results are sightly worse for longer horizons, consistently with what was shown in a related study [119].

### 7.5.4 Qualitative results

Figure 7.9 shows the most probable deformation when sampling the latent space for each phase spanning one respiratory cycle, as well as the ground-truth motion fields. It can be observed the spatio-temporal consistency of the predictions with respect to the ground-truth deformation. Knowing that every sampling yields a new result, an uncertainty map can be constructed by sampling $N$ times. This represents a measure of the statistical dispersion of each attributed value and therefore gives an idea of the areas more prone to errors. Such functionality is possible due to the probabilistic and generative nature of the model. Figure 7.10 displays the uncertainty over the whole organ across several phases.

Figure 7.11 shows qualitative results on the V-MRI dataset at several respiratory phases. In each image, the left half belongs to the ground truth and the right half belongs to the prediction. Dashed lines indicate the maximum motion amplitude from end-exhale to end-

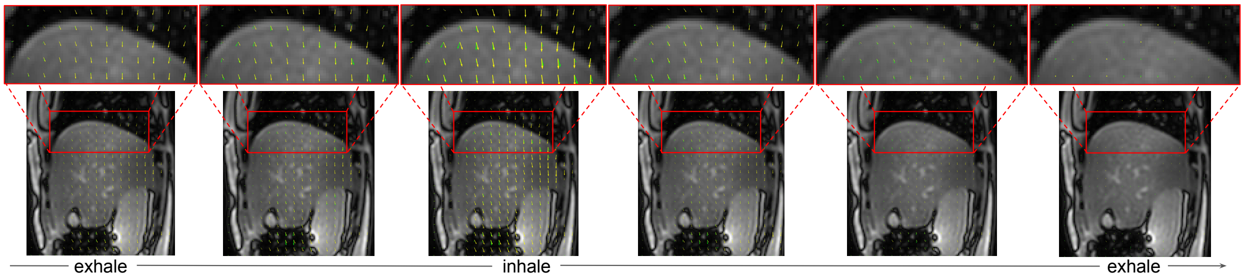


Figure 7.9 Most probable deformation fields chosen when sampling the probabilistic latent space for several phases spanning one respiratory cycle. Green and yellow arrows represent ground-truth and predicted motion fields, respectively.

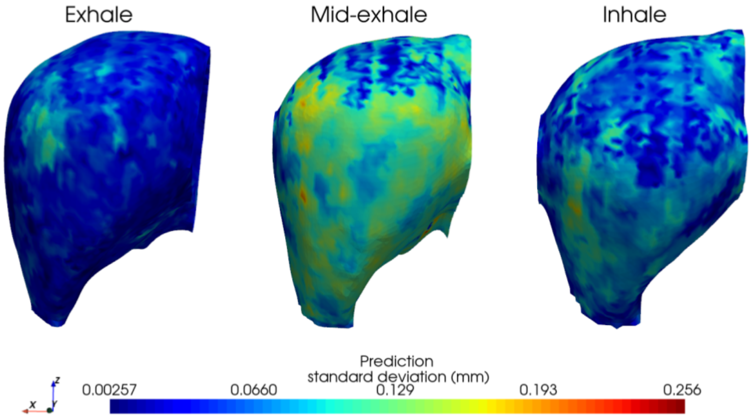


Figure 7.10 Motion-based prediction uncertainty maps ($N = 50$) at selected respiratory phases.

inhale.

Figure 7.12 presents the outputs of the proposed model, comparative approaches as well as the true volume for different respiratory phases with the US dataset. The outputs are presented in the sagittal and axial planes to demonstrate that motion is generated in all three dimensions. It can be observed that difference maps corresponding to ME present some regions with high error values. Although the results of the other two comparative approaches (DN and AE) look similar to those showcased by the proposed model, there are some small areas, particularly within the liver, where values are slightly worse for these methods. Red circles are added to point out differences in the generation quality of anatomical features like vessels and liver borders. In both cases, comparing ground truth and predictions, it is noticeable that the model correctly predicts the motion shown by the true image sequence. Additional qualitative results can be found in the Supplementary materials.

The proposed method requires a computational time of 10.2 ± 0.6 ms when deployed on a NVIDIA Titan RTX GPU with 64 Gb RAM to predict 3 future deformation fields on the V-MRI dataset, which is equivalent to a horizon of 1.3 s. This value was obtained by averaging 50 different inference times. According to the literature, predicting between 300 and 600 ms ahead of time is enough to cope with the typical system's latency during radiation delivery [174, 203]. Hence, considering its low computational time, the motion model is sufficiently performant to be applicable in real-time and to allow precise online tracking of the target volume. Each training in the V-MRI dataset took approximately 4 hours. The time required to perform a subject-specific fine-tuning was 8 min.

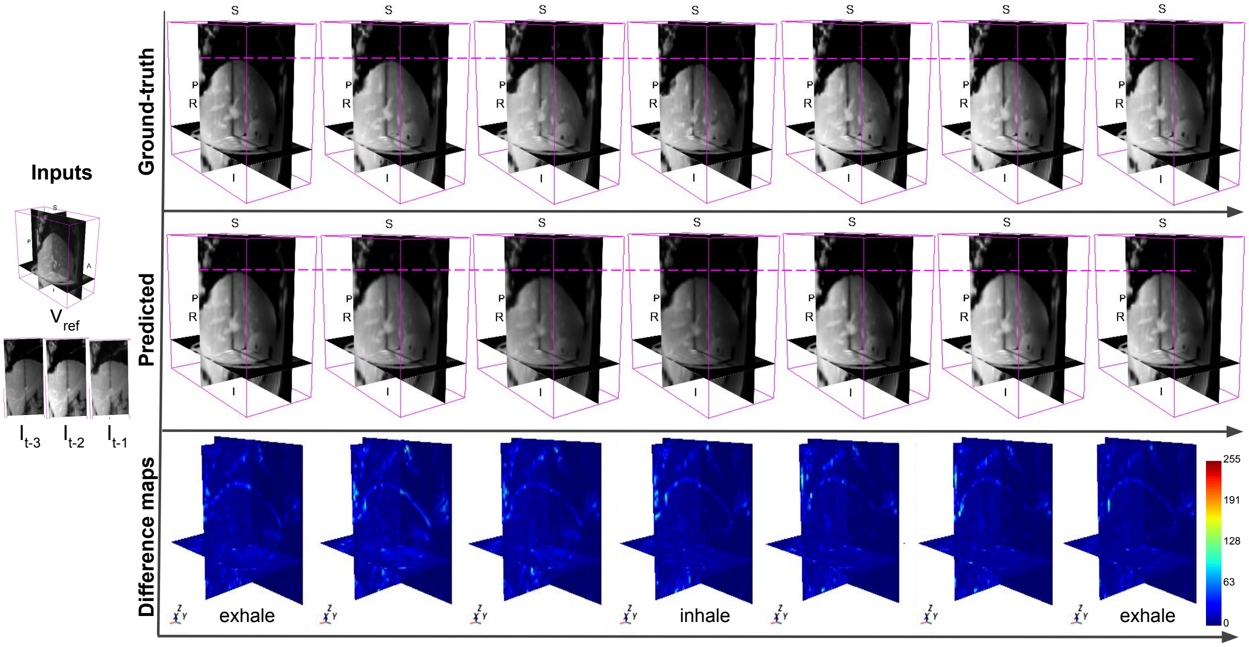


Figure 7.11 Qualitative results and difference maps between ground-truth and predicted volumes.

## 7.6 Discussion and Conclusion

We presented an unsupervised predictive framework that can generate 4D volumes given only a reference pre-treatment volume and real-time 2D slices. Our method not only allows accurate spatio-temporal predictions with 1.3s time horizon, but also provides uncertainty values. Indeed, deep neural networks are known by their capacity to approximate a large class of functions. The proposed generative framework, inspired by a conditional variational autoencoder, relates partial observations to dense 3D motion fields over time. These partial observations are in-room cine slices or US images that capture the internal organ motion. This has an advantage over models relying on external surrogates that are not always representative of the actual organ motion.

During the model validation, we considered imaging modalities such as MRI and US, which avoid an extra burden of ionizing radiation for the patient. Unlike previous approaches that have constructed motion models relying on treatment planning 4D CT or CBCT, the two datasets employed in this study were acquired during free-breathing. Generally, the treatment planning data is a respiratory-correlated dataset which only captures an average motion over the respiratory cycle. This may yield poor motion estimation results and difficulty in coping with irregular patterns. In contrast, our time-resolved free-breathing predictive model covers

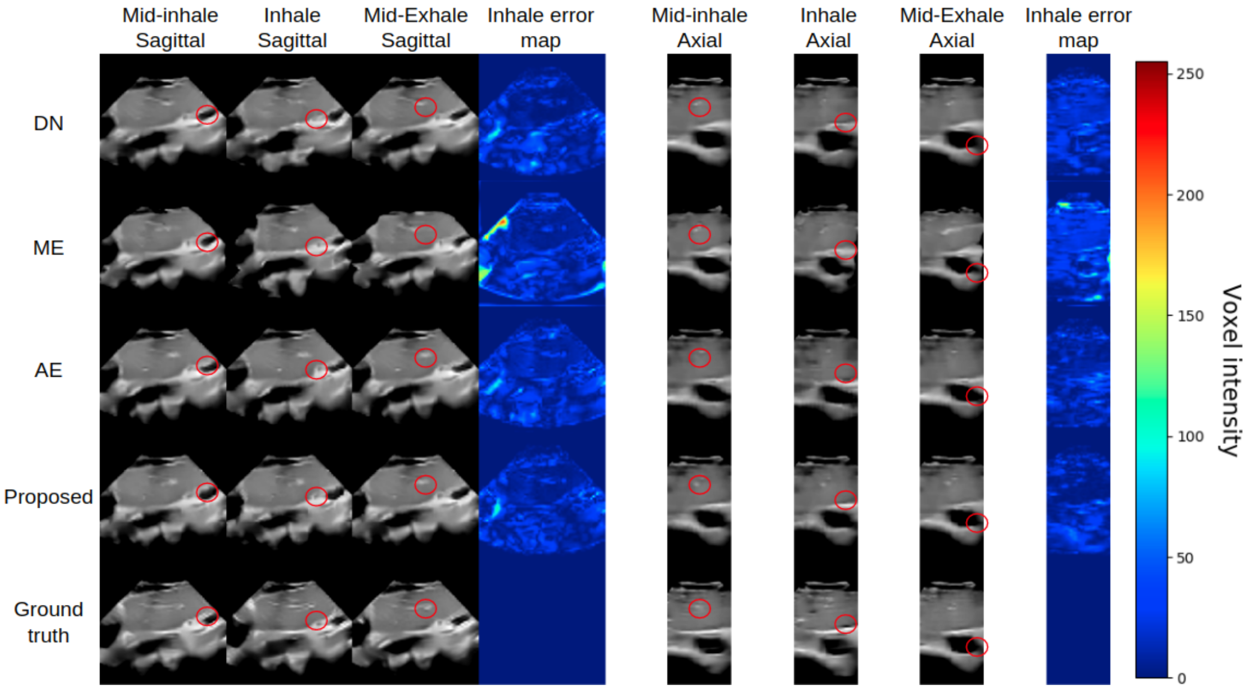


Figure 7.12 Qualitative results for all compared methods on the US dataset. For both sagittal and axial planes, the central slice of the volume is shown at mid-inhale, end-inhale and mid-exhale respiratory phases. For the end-inhale lisetpoly phase, an error map is calculated. Red circles are included to highlight differences between the displayed approaches.

a wide variability of breathing patterns and the model showed to be robust managing cycle-to-cycle variations. Furthermore, it can accommodate to challenges which may occur during therapy. In this regard, a different image acquisition protocol, liver drift motion over longer time and coverage of the whole liver were taken into account in our experiments. The hold-out dataset used for testing purposes was acquired and reconstructed with a different MRI protocol to the one used to build the motion model. This experiment is crucial to validate whether or not the model can work with different image contrasts and appearances, which is very likely to happen in the clinical routine. In terms of respiratory organ drift, related works have shown how the performance decreases up to four times between the extreme subsets [244]. In contrast, the proposed model shows little quality degradation with time thereby offering superior robustness compared to existing techniques.

The analysis of the motion embeddings revealed that the latent space projects similar respiratory phases close to each other. This could be potentially used for further classification. The ablation experiments showed the influence of each component involved in the model, demonstrating their effect on the overall performance. It was observed that considering the temporal consistency of the slices concatenated with the corresponding slice in the reference volume and adding anatomical features is beneficial for the performance. Moreover, processing the image sequence with ConvLSTM proved to be the most effective way to enable multi-time predictions, which is consistent with previous results [119]. The proposed model enables one-shot multi-time predictions. Results were reported when the number of future time steps ranged from one to three. Nonetheless, it is important to note that the optimal horizon will depend mainly on the clinical application.

In the comparative study, we observed significant differences between the results obtained with sagittal and coronal orientations in the MRI datasets. In the US data, axial orientation outperformed the other alternatives. Presumably, more anatomical information with larger fields of views favors the predictions, which is in line with results reported in previous works [90, 190, 233]. The proposed model tries to address some limitations of the existing solutions. For instance, the main shortcoming of simplified strategies, such as the spatial motion extrapolation [202], is related to the derivation of an accurate global anatomical description since it was designed for local modeling. Also it lacks of a temporal predictive mechanism. In contrast, our model enables dense spatio-temporal predictions from the same inputs, i.e. a 3D pre-beam volume and beam-on cine images.

Previous studies have described the hysteresis trajectory followed by the liver during free-breathing, which is visible in the SI and AP planes [2, 206]. Therefore, intuitively one may think that using coronal slices might affect the hysteresis recovery, since this orientation does not capture one of these planes. However, because predictions are based on breathing phase-detection rather than image appearance, even if the orientation of the surrogate image does not capture the SI-AP motion, the predicted DVF does contain the hysteresis. For instance, Figure 7.6 showcases SI and AP trajectories yielded from coronal slices. In fact, the hysteresis is reflected in the training motion fields, and subsequently learned by the motion model. Since the respiratory phases are linked to high-dimensional deformations, by recovering the phase computed from the surrogate (regardless of the orientation) the model will try to generate the corresponding motion fields similarly as learned during training.

When comparing the tracking capabilities, the approach based on merging features from a reference volume and surrogates yielded the poorest performance. This showcases that the proposed model benefits from a manifold-structured latent space, as previously illustrated in Figure 7.5. As an added value, such representation fosters the model's interpretability. Furthermore, decision-making in the low-dimensional space has advantages over current surrogate-driven statistical models. In these cases, the weights optimization relies on a similarity metric based on high-dimensional information, which only captures the variation in a single plane. Consequently, the surrogate used to update the model does not consider a global adaptation. Conversely, in our approach, with the same type of input we tackle this shortcoming by working with latent representations, which contains 3D information thereby ensuring global adaptation. Moreover, the limited capacity in properly compensating for motion in a third dimension when using a single plane cine-MRI has been acknowledged in previous studies [233]. This limitation is generally resolved by combining sagittal and coronal slices at the expense of increased computational cost, though. On the other hand, in statistical models, the fixed relationship assumed between the motion of the surrogate itself and that of the entire anatomy is known to be a limitation, especially under irregular breathing conditions [233]. As consequence, these models perform well only on patients showing little differences between the mean cycle signal and the actual free-breathing signal, i.e., in regular cycles, similar to the average motion seen during model construction. In these cases, the similarity function used for weight optimization quickly converges to the optimal solution. In contrast, our model showed enough capacity to cope with irregularities in the respiratory cycle. While some individuals present stable and regular breathing cycles, others have irregular patterns (coughing, sneezing) that can lead to the internal target volume underestimating the true range of motion [245]. The model also demonstrated certain tolerance to potential shifts of the surrogate position, which is an important characteristic especially for ultrasound, where it is more difficult to reproduce a certain imaging plane.

Unlike a related autoencoder-based model [190] trained in 3 steps, our approach only involves a single training step. The proposed framework is flexible in terms of imaging modality and pre-treatment data availability. Moreover, it is neither limited to a specific organ, nor to radiation therapy as a treatment modality. In fact, other applications requiring motion compensation can be considered. In theory, any type of images could be used as surrogate independently of the output imaging modality. The only aspect that would need adapting is the synchronized acquisition of both datasets, similarly as [206] and [180]. Furthermore, the described method can be applied both as population-based or as subject-specific. If we consider the first scenario, an important advantage of our model is that, due to the strong generalization capabilities of the neural networks, it does not require finding inter-subject correspondences. This process, which is part of the workflow in statistical modeling, is time-consuming and needs manual interaction. In the second scenario (subject-specific mode), the learning process can benefit from a broad motion and anatomical variability before the personalization to a given patient. Whenever possible, this scenario is preferable since it yields more accurate predictions. Recent approaches using partial Fourier acquisition, different k-space read-out strategies and the use of deep learning have shown promising results to shorten reconstruction times [64,218]. Therefore, this could become a viable option to quickly obtain in-room 4D MRI datasets to personalize the population models before treatment.

While some current clinical systems enable target monitoring using 2D orthogonal images, the unobserved intra-fraction motion of the organs-at-risk may degrades dosimetric benefits [215]. This issue can be alleviated by proving volumetric information. Hence, the predicted dense displacements fields can be used as a feedback variable for dose calculations, real-time online plan adaptation and anatomy tracking during interventions. The level of accuracy achieved in this work (1.67 mm and 2.17 mm for P-MRI and US datasets) is deemed sufficient according to the standards mentioned in related works [22]. For instance, the population model proposed by [206] achieved a mean error of 2.4 mm over 8 subjects. This model was intended to be used for proton therapy. Furthermore, the authors mentioned 3 mm as a clinically acceptable margin accuracy for this treatment modality. Similarly, the Lung Target Tracking Challenge [246] considered tracking errors lower than 2 mm as a clinically relevant primary ranking metric.

With regard to the dataset, it is worth mentioning that errors stemming from the reconstruction process (e.g. discontinuous organ edges between consecutive slices) may negatively affect the image registration process. Therefore, the quality of the employed dataset must be ensured to yield plausible motion fields. Likewise, it is important to consider that similarly to other motion models such as PCA-based, the performance will depend on the accuracy of the deformable registration, which provides the training data for model creation. Certainly, deformable registration is a process that is not exempt from errors. Moreover, in practice, there are no ground-truth motion fields available. In our experiments, the estimated average error after registration between the moving and fixed volumes was $1.0 \pm 0.6$ mm. We consider this value as an acceptable threshold accuracy for the modeling task. Also, some considerations should be made regarding the acquisition of the reference volume. In the current experiments, it was extracted from the 4D dataset. Nevertheless, in the clinical scenario this reference will be a breath-hold image acquired before therapy. This difference should not represent an obstacle as long as the field-of-view is consistent with the one used for the training volumes and all the images are aligned to a common reference system. These requirements can be met with a proper setting of the scanner. Finally, in terms of computational cost, training and/or fine-tuning is performed before treatment whereas during treatment, only inference is required. The inference occurs within a few milliseconds, which allows the radiation device to react on the estimated target motion. Future studies should focus on evaluating the model performance from a dosimetric point-of-view.

**Acknowledgments** This work was partly funded by an NSERC collaborative research and development project (CRDPJ-517413-17) and by the Canada First Research Excellence Fund through the TransMedTech Institute.

# CHAPTER 8 ARTICLE 4: PREDICTION OF IN-PLANE ORGAN DEFORMATION DURING FREE-BREATHING RADIOTHERAPY VIA DISCRIMINATIVE SPATIAL TRANSFORMER NETWORKS

Contribution of the first author in preparation and writing this paper is evaluated as 90%. This article has been published by Medical Image Analysis journal on June 2020.

**Remarks:** This paper presents a recurrent encoder-decoder architecture which leverages feature representations at multiple scales. It simultaneously learns to map in-plane deformations between consecutive images and to extrapolate them through time. Experimental results on healthy subjects and patients revealed that this approach yields a clinically relevant accuracy while presenting important advantages over similar state-of-the-art methods.

## Prediction of in-plane organ deformation during free-breathing radiotherapy via discriminative spatial transformer networks

Liset Vázquez Romaguera[1], Rosalie Plantefève[2], Francisco Perdigón Romero[1], François Hébert[3], Jean-François Carrier[4], Samuel Kadoury[1,2]

[1] École Polytechnique de Montréal, Montréal, Canada [2] Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Montréal, Canada, [3] Elekta Ltd., Montréal, Canada, [4] Centre Hospitalier de l'Université de Montréal and Département de physique, Université de Montréal, Montréal, Canada

### Abstract

External beam radiotherapy is a commonly used treatment option for patients with cancer in the thoracic and abdominal regions. However, respiratory motion constitutes a major limitation during the intervention. It may stray the pre-defined target and trajectories determined during planning from the actual anatomy. We propose a novel framework to predict the in-plane organ motion. We introduce a recurrent encoder-decoder architecture which leverages feature representations at multiple scales. It simultaneously learns to map dense deformations between consecutive images from a given input sequence and to extrapolate them through time. Subsequently, several cascade-arranged spatial transformers use the predicted deformation fields to generate a future image sequence. We propose the use of a composite loss function which minimizes the difference between ground-truth and predicted images while maintaining smooth deformations. Our model is trained end-to-end in an unsupervised manner, thus it does not require additional information beyond image data. Moreover, no pre-processing steps such as segmentation or registration are needed. We report results on 85 different cases (healthy subjects and patients) belonging to multiples datasets across different imaging modalities. Experiments were aimed at investigating the importance of the proposed multi-scale architecture design and the effect of increasing the number of predicted frames on the overall accuracy of the model. The proposed model was able to predict vessel positions in the next temporal image with a median accuracy of 0.45 (0.55) mm, 0.45 (0.74) mm and 0.28 (0.58) mm in MRI, US and CT datasets, respectively. The obtained results show the strong potential of the model by achieving accurate matching between the predicted and target images on several imaging modalities.

**Keywords:** Motion prediction, free-breathing, liver, lungs, radiotherapy, deep learning, LSTM

### 8.1 Introduction

External beam radiotherapy (EBRT) is a commonly used treatment option for patients with cancer in the thoracic and abdominal regions, for example, in lungs and liver. Statistics reveal that between 40 - 50% of patients diagnosed with this disease undergo this treatment [247]. During the procedure the goal is to optimize the dose over the tumor while sparing healthy tissue. However, there are several factors contributing to potential inaccuracies during the treatment. Among them, respiratory motion induced by free-breathing is one of the major issues in abdominothoracic radiation treatment and have shown to have a large dosimetric impact [21]. Organs such as the lungs, liver, pancreas and kidneys are known to move dramatically with respiration. Several studies, particularly in hepatic imaging, have shown the extent of various modes of liver deformation during free-breathing [169, 182]. The main motion component in the liver has been measured in the superior-inferior direction, with a typical range of 5-25 mm for relaxed breathing, whereas in anterior-posterior (1-12 mm) and left-right (1-3 mm) directions the motion amplitudes are smaller. Moreover, these studies have demonstrated that the nature of deformation is much more complex than a simple caudal-cranial translation, and includes elastic deformation as well as rotation effects which might affect the dose administration towards a defined target [248, 249]. On the other hand, although breathing shows a repetitive pattern, there is an inter-cycle variability that is not negligible. This can be evidenced during involuntarily shallow or deep breathing which represent an additional challenge during treatment.

Several solutions have been proposed to deal with the problem of respiratory motion during imaging and image-guided interventions such as EBRT. According to [45], they can be classified into two categories: non-adaptive motion compensation and real-time adaptive methods. The first category includes techniques such as using large Planning Treatment Volume (PTV) margins, abdominal compression, treating during breath hold, or respiratory gating [22]. Using large margins to cover the whole range of tumor motion is clearly undesirable as this increases the volume of healthy tissue exposed to high doses of radiation. Forced shallow breathing using a stereotactic body frame is one alternative to use during radiation treatments. It reduces diaphragmatic excursions, while still permitting limited normal respiration [46]. During treatment, the images are essential to verify the position of the tumor due to the difficulty in reproducible placing the abdominal compression device. Breath-holding is a simple approach but it limits acquisition/intervention time to typically less than 30 seconds. Moreover, some patients may not be able to tolerate this procedure. Respiratory gating involves only acquisition/treatment during a limited portion of the respiratory cycle (e.g. end-exhalation). However, it significantly increases the procedure time [22].

On the other hand, adaptive motion tracking re-positions the radiation beam as the tumor moves. In this approach, organ motion modeling is a crucial component. We can distinguish two types of motion modeling. Local approaches use information surrounding the target to reconstruct exclusively the 2D/3D position of the tumor, while global approaches relying on in-room surrogates and patient-specific global motion models estimate the whole 2D/3D anatomy [49]. Some of the clinically available solutions follow the local approach. For instance, the CyberKnife Synchrony system [250] relies on the construction of a correspondence motion model between respiratory surrogate signals and the tumor motion. Surrogates are acquired by measuring the displacement of the patient's abdomen or chest using optical devices [45]. As it can be difficult to image the internal motion, fiducial markers such as gold seeds are often implanted percutaneously near the region of interest and tracked using fluoroscopy. Such implantation techniques are invasive and motion information is only available at the marker(s) and not for the whole region of interest.

Over the last few years, Magnetic Resonance Imaging (MRI) have emerged as an image guidance modality in radiotherapy (RT) treatment units, thereby creating an entirely novel paradigm denoted as MR-guided RT (MRgRT) [223]. Moreover, technological innovations in dose delivery systems, such as the MR-Linac, have enabled the acquisition of high-quality, real-time images immediately before, during and after the patient is treated [55,223,251,252]. Similarly as EBRT, other tumor ablation modalities benefit from image guidance: radiofrequency ablation is normally guided using ultrasound (US) images whereas high-intensity focused ultrasound (HIFU) can be guided using MRI [224, 225] as well as diagnostic US data [49, 226, 227]. One major limitation of image-guided interventions is the acquisition of sparse 2D slices, which does not guarantee an accurate adaptation for 3D motion [4]. Consequently, there has been several attempts to derive 3D motion from partial observations (single or interleaved slices), demonstrating promising results [90,179,202,204]. For instance, a quantitative comparison between five established strategies that derive time-resolved volumetric MRI in MRI-guided radiotherapy was presented by [233].

In order to cope with system latencies between target localization and dose delivery, the integration of the aforementioned works with in-plane sequential prediction represents a potential and feasible short-term alternative to produce real-time 4D. Therefore, the in-plane motion prediction is a relevant task which we address in the present work.

### 8.1.1 Related works

A vast amount of literature has presented various two dimensional (2D) and three dimensional (3D) predictive deformation models. Generally, these methods rely on statistical modeling [13, 90, 178, 179, 235, 253, 254], biomechanical modeling [185, 255, 256], atlas creation [182], clustering [257], template matching [258] and deformable image registration [48, 202, 259]. Among these approaches, statistical modeling is one of the most explored and has proven to achieve state-of-the-art results. Hereinafter we refer to all of them as model-based approaches. Although most have been used for 3D prediction, the method itself is also applicable for the 2D case.

**Model-based approaches:** Principal Component Analysis (PCA) is a well-established statistical approach that has been used for the construction of subject-specific and population-based motion models. Generally, eigen decomposition is performed on a motion matrix which is obtained from deformable image registration (DIR) between a reference phase and other phases in a four-dimensional (4D) dataset. According to [253], every possible organ motion state can be approximated by a linear combination of the eigenvectors corresponding to the largest eigenvalues. Previous studies have found that two principal components are adequate to describe respiratory motion in 4D-CT datasets [170, 253]. However, PCA models by themselves can only provide spatial predictions without any temporal consideration. Consequently, several works have proposed to integrate temporal prediction using surrogates with the model's spatial prediction. [235] built exemplar models by fitting a PCA model to the motion vectors of each individual subject. The final model was a weighted combination of the predictions from all the sub-models. In a follow-up study, liver vessels landmarks were tracked and temporally extrapolated to achieve spatio-temporal prediction from the PCA model [184]. Linear adaptive filter was chosen for temporal prediction since it showed the best performance in a comparative analysis presented in [260]. [254] combined a statistical model and information from 2D ultrasound sequences. They used an artificial neural network with a single hidden layer for temporal prediction on anatomical landmarks tracked in the ultrasound images. Similarly, [90] parameterized 3D motion information using PCA. The weights of the eigenvectors were iteratively optimized until the warped reference volume matched the incoming interleaved slices. Inspired by [178], [179] proposed a model based on anatomical regions of interest to relate the 3D motion, derived from 4D-CT data, with CT slices centered around the tumor.

Most of existing 2D motion prediction methods follow a local approach or utilize low-dimensional position information to derive the predictions. For instance, Yun et al. (2012) implemented an artificial neural network to predict lung tumor positions in cine-MRI images. Similarly, Bourque et al. (2017) proposed a 2D motion prediction algorithm for lung tumors using a particle filter combined with an autoregressive model. This approach enabled to sequentially track and predict the tumor position, 250 ms in the future. Both works were developed in the context of MR-linac treatments. Seregni et al. (2016) evaluated different predictive algorithms, namely, linear extrapolation, autoregressive model and support vector machine. The tumor positions were identified in cine-MRI slices using scale invariant features. The authors showed that cine-MRI guidance, combined with prediction algorithms, could decrease geometric uncertainties during treatment. Alternatively, Ginn et al. (2020) leveraged the high-dimensional information carried by the images to drive the predictions. Specifically, they used a weighted combination of previously observed motion states. The weights were determined by calculating the sum of squared differences (SSD) between the current and past images. However, SSD is known to have poor performance when noise corrupts the image intensities. Another limitation is that this approach may not provide accurate predictions for irregular motion not captured in the selected most similar images.

In the past few years, deep learning techniques have been proposed for spatio-temporal motion prediction on image sequences. However, most of the works have been reported in the field of natural images while in medical imaging the contributions are scarce. These data-driven approaches automatically discover and learn discriminatory features from a very large number of labelled or unlabelled examples. That information is then used to perform a prediction task such as regression or classification [261].

**Learning-based approaches:** Several deep learning techniques have been proposed for motion prediction on natural images [9], as well as video frame prediction [8,101–103]. Moreover, different models were proposed in complex scenes to predict the actions [262], poses [263] or trajectories of humans [264]. Some of these architectures have been inspired from the recurrent encoder-decoder model developed by [148] for machine translation. In this groundbreaking work, the encoder is composed by a multilayered Long Short-Term Memory (LSTM) that maps the input sequence to an internal representation called "context vector". Subsequently, another LSTM (decoder) is used to generate the output sequence from the vector. The lengths of the input and output sequences can be different, as there is no explicit one-to-one relation between the input and output sequences. Such idea was further extended by [101], who developed a LSTM encoder-decoder framework for reconstruction and future frame prediction. In [265], the authors introduced a generative model that uses a recurrent neural network to predict the next frame or interpolate between frames. [120] proposed a convolutional neural network to predict dense optical flow given a static image. The authors stated that motion estimation via regression has an important drawback: the output space tends to smooth results to the mean. In consequence, they turned the problem into a classification task. Similarly, [9] posed the motion prediction on natural images as a classification task and hypothesized that directly regressing the displacement fields is not a suitable approach.

Generative models have received a considerable attention for future frame prediction [145, 162, 165, 266]. However, these approaches suffer from blurriness which is likely due to the difficulty in directly regressing to pixel values [166]. This limits its application, particularly in the medical field. An alternative approach is to learn a disentangled representation from image sequences. For instance, [135] and [8] factorized the frames into a stationary part and a temporally varying component to represent content and motion, respectively. The prediction of future dynamic is then enabled applying standard LSTM to the time-vary components. Although the spatio-temporal dependency in sequential data has been well explored and many recurrent variants have been considered in deep learning for motion prediction, the results presented in the vast majority of cases are based on validations carried out in datasets with basic or low complexity movements (e.g. KTH action dataset [267], KITTI [268], Waving Flags [103]), or even in synthetic or test data sets (e.g. Moving MNIST [101]).

On the other hand, relatively little has been done on this matter in the medical imaging field. Recently, [269] proposed a conditional variational auto-encoder that can predict motion from an image sequence. However, the model was only validated on cardiac MRI-cine data and requires complete cardiac cycles as inputs. Similar works have addressed the motion estimation on cardiac sequences using siamese-style convolutional recurrent units [270] and manifold learning theory [271]. Several configurations using LSTM have been proposed to address related predictive tasks [272–277]. Specifically, for application in radiotherapy, most of the works have been focused on temporal prediction from the Real-Time Position Management [278] system acquisitions [272, 273, 275, 277] while scarce works have been presented on future frame generation. This work is motivated by the aforementioned challenges. Therefore, we aim to propose a novel mechanism for future frame generation and to validate it with clinically relevant cases across the most commonly used medical imaging modalities.

### 8.1.2 Contributions

In this work, we introduce a novel recurrent encoder-decoder architecture to perform multi-time in-plane motion prediction. We leverage feature representations at multiple scales and convolutional LSTM to find the deformation between input images and to learn how to extrapolate them through time. Then, spatial transformers take over the image generation process. The main contributions can be summarized as follow:

- We propose a novel multi-scale recurrent encoder-decoder model for motion prediction in multiple times (Section 8.2.2). We introduce a differentiable spatial transformation for displacement fields (implicit) regression and image generation (Section 8.2.3).
- We evaluate the modeling power of our method across different imaging modalities and we show that our pipeline outperforms state-of-the-art approaches (Section 8.3).
- We show that classification-based models can greatly benefit from an adaptive motion encoding (Section 8.3.4).

We postulate that the introduction of deep learning-based motion models will become an important component toward the next generation of image-guided radiotherapy. Moreover, the development of a new family of motion compensation methods which will impact directly tumor targeting during radiotherapy.

## 8.2 Motion prediction network

In this section, we present our in-plane motion prediction pipeline. As highlighted in the previous section, our approach combines a multi-scale recurrent encoder-decoder model with

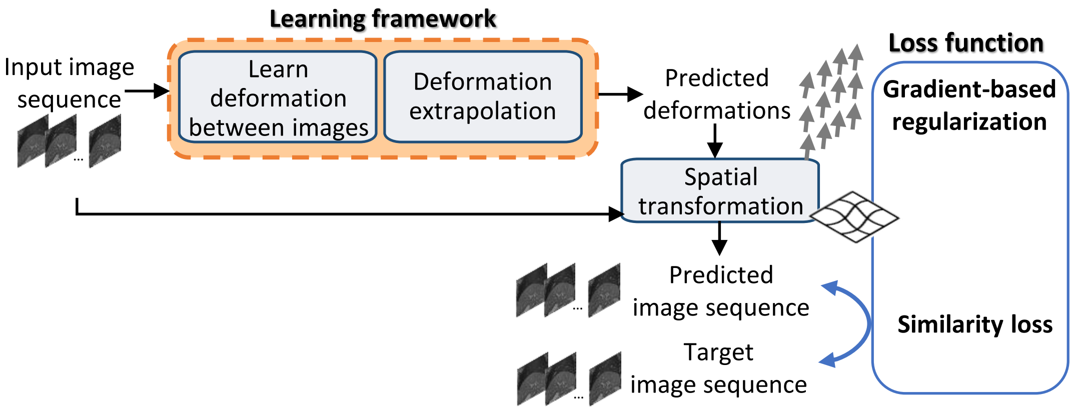


Figure 8.1 Functional blocks of the motion prediction framework. The model first learns the successive deformations in the input sequence. Then the feature representation is extrapolated and further upsampled up to the original dimensions. The spatial transformer warps the input image with the predicted deformations in order to generate the future frames. A composite loss function takes over of minimizing the differences between predicted and target sequences as well as ensuring smooth deformations.

a differentiable spatial transformer module. The goal is to implicitly regress the future spatiotemporal 2D deformations for image sequence generation. Figure 8.1 shows the functional blocks of the proposed in-plane motion prediction framework. First, a neural network learns how to align the images. In other words, it learns to non-rigidly register image pairs. Secondly, the feature representation corresponding to the extracted deformations is extrapolated and further upsampled up to the original dimensions. Then, the last input image is warped with the predicted deformation in order to generate future images. Finally, a composite loss function takes over of minimizing the differences between predicted and target sequences as well as ensuring smooth deformations.

In the following sections, we state the task in question (Section 8.2.1), describe the proposed motion learning framework (Section 8.2.2) and the spatial transformer module (Section 8.2.3), as well as the loss function used to train our model (Section 8.2.4). Finally, we provide details about the training protocol (Section 8.2.5).

### 8.2.1 Problem formulation

For the spatio-temporal in-plane motion prediction problem, we consider an ensemble $D$ of sequentially-acquired population data (volunteers or patients), namely 2D + time datasets. The motion in each dataset $d \in D$ can be quantified by performing pair-wise deformable registrations between consecutive pairs of images. Therefore, each dataset can be represented as a sequence of 2D motion fields which contains the vectorial components that express the deformable displacement of the organ between two given times $t_k$ and $t_{k+1}$.

The proposed model aims at learning a representation that predicts the sequence of dense displacement fields which represents the deformation of a given organ during free-breathing acquisitions. We formulate this task as follows: given a temporal input $\left(i\right)$ image sequence $\mathbf{X} = \left\langle I_1^i, I_2^i, \ldots, I_n^i \right\rangle$ of length $n$, our goal is to predict the sequence of motion fields $\mathbf{\Phi} = \left\langle \phi_n, \phi_{n+1}, \ldots, \phi_{n+T} \right\rangle$ over $T$ time steps, where $\phi_n$ is the predicted motion at time $n$. Moreover, $\mathbf{\Phi}$ contains the deformations corresponding to a future output $\left(o\right)$ image sequence $\mathbf{Y} = \left\langle I_{n+1}^o, I_{n+2}^o, \ldots, I_{n+T+1}^o \right\rangle$ with the same length $T$; $I_{n+1}^o$ results from warping the previous image $I_n^i$ with $\phi_n$; $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{H \times W}$, $\mathbf{\Phi} \in \mathbb{R}^{H \times W \times 2}$ where $H$ and $W$ are the height and width of the images, respectively. Therefore, the proposed model aims to learn from the training data a discriminative model by means of a $d$-dimensional latent distribution that maximizes the conditional probability $\mathrm{P}(\mathbf{Y}|\mathbf{X})$ of obtaining a predicted sequence given the input sequence.

### 8.2.2 Motion learning architecture

The outline of the proposed in-plane motion prediction architecture is shown in Figure 8.2. It is composed by a fully convolutional spatial encoder, recurrent units, a fully convolutional spatial decoder and multiple spatial transformer layers, depending on the number of predicted time steps. The length of the input and output sequences can be empirically determined, depending on the nature of the application. The spatial encoder is fed with an input temporal sequence. It extracts high-level features from the input images and maps the consecutive deformations between them. Afterwards, the spatio-temporal features are extrapolated in time by the recurrent units, which are then processed by the spatial decoder to recover the desired dimensions in the form of smooth deformations. Finally, the deformations are applied in cascade through individual spatial transformer layers yielding the predicted image sequence.

Previous works [9, 118, 120] on motion prediction and missing frame interpolation have addressed the feature extraction problem by using repeated convolution and pooling layers. With this traditional architecture, fine-scale features are learned in the early layers whereas coarse-scale features are learned in later layers. However, there is no guarantee that a certain feature, which may be crucial for the performance, will be maintained through the contracting paths until its extraction in deeper layers. Therefore, instead of using the aforementioned approach, we introduce an alternative scheme that extracts feature representations at multiple scales through the network. Figure 8.3 illustrates the internal structure of the proposed multi-scale residual (MSR) block which processes the input tensor at different levels: full resolution, medium resolution and low resolution in order to fully exploit the image features.

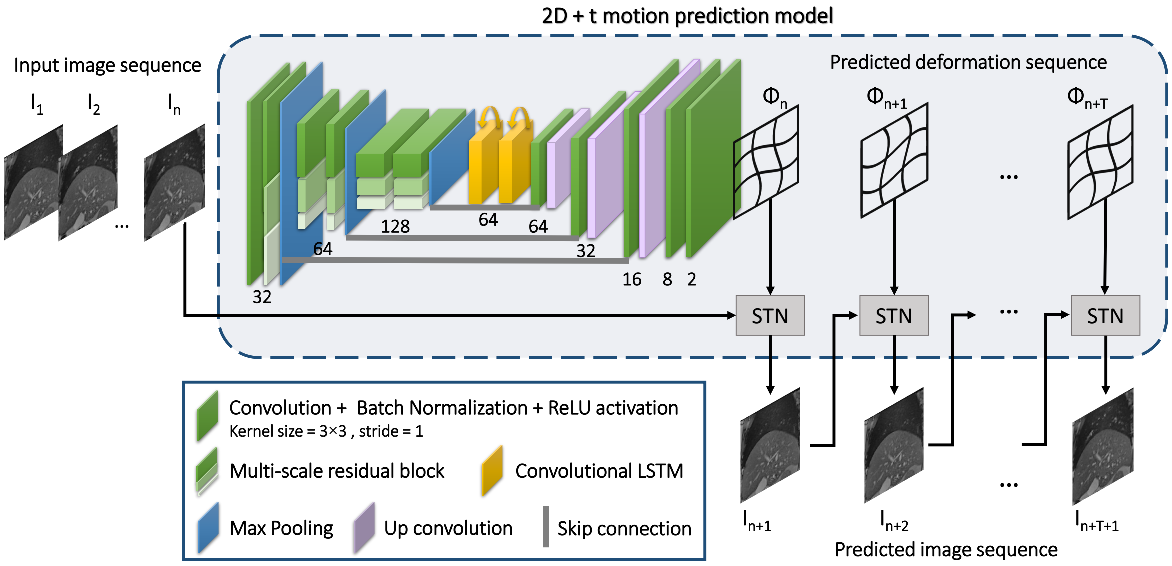


Figure 8.2 Schematic representation of the proposed model. The model receives a image sequence as input and learns to generate future frames. Convolutional layers, arranged in a multi-scale configuration, extract features at different resolutions and map the consecutive deformations between the input images. Convolutional LSTMs leverage the spatio-temporal properties of the sequence and extrapolate the feature representations which are further upsampled by up-convolutional layers in the decoder. Interleaved convolutional layers decrease the features dimension up to a plane representing the anterior posterior and superior inferior motion directions. Finally, cascade-arranged spatial transformers apply the predicted deformations to generate the output image sequence.

In the first pathway, features are extracted from the input without reducing its size. In this manner, fine-grained features are extracted from the images in original size since there is no pooling layers. In the other two pathways, the input resolution is decreased up to two times by using average pooling layers. The reason behind this choice is that, while max pooling only retains a quarter of the data, average pooling brings all into account retaining more information.

In the two coarser levels of the block, the input of the convolutional layers is enriched with features extracted in the convolution of the previous scale. To this end, transversal max pooling modifies the tensor size to match with the current level. Both tensors are concatenated and fed to the convolutional layer. With this configuration, features from the highest levels are shared with lower levels across the block. The core idea is to provide the network with multiple ways to extract information pertaining to deformation. Up-convolutions upsample the feature maps to the original block input size. At the end of each pathway, features extracted at different scales are merged in the concatenation block. Inspired by ResNet [279], we placed a shortcut that performs identity mapping on the input and add it to the MSR block output. All convolutional layers have $3 \times 3$ receptive fields, single stride and ReLU activations. Batch normalization layers were placed after convolutions as it speeds up training and acts as a regularizer.

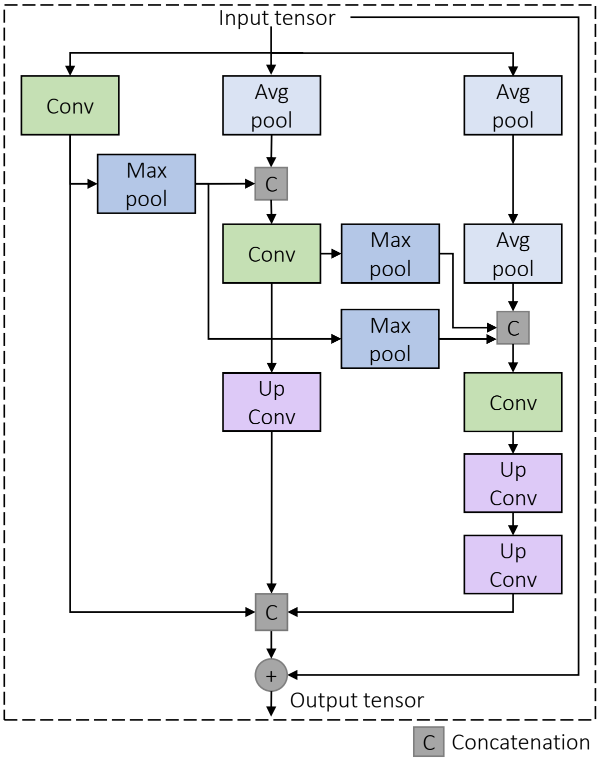


Figure 8.3 Internal structure of the proposed multi-scale residual block. The input tensor passes through three pathways consisting of convolutional and pooling layers. This scheme allows for feature extraction at multiple resolutions. At the bottom part, features are upsampled to the original dimension and combined to shape the output tensor.

In order to perform prediction, Vanilla LSTM have shown excellent results in processing sequences of one-dimensional vectors. However, its performance is limited in handling multi-dimensional sequential data such as images (2D + time) and volumes (3D + time) due to lacking spatial correlations. We leverage the variant proposed by [112] in which internal point-wise matrix multiplications are replaced by convolutional structures for both input-to-state and state-to-state transitions. By doing so, it captures underlying spatial features using convolution operations in multiple-dimensional data. These cells, named convolutional LSTM (ConvLSTM), are able to capture spatio-temporal properties of the data much better than vanilla LSTM cells and have shown to outperform them with even containing fewer model parameters. These memory units keep an accumulative knowledge across sequences through their cell state $C$. Internal gates provide these units with the capacity of erasing information that is no longer useful (forget gate $f$) and adding relevant information (input gate $i$). Such decisions are made by neural networks with sigmoid activation functions $\sigma(\cdot)$, weights $W$ and bias $b$. At each iteration the old cell state $C_{t-1}$ is updated into the new cell state $C_t$:

$$C_t = f_t \circ C_{t-1} + i_t \circ tanh\Big(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c\Big) \tag{8.1}$$

$$f_t = \sigma\Big(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \circ C_{t-1} + b_f\Big) \tag{8.2}$$

$$i_t = \sigma\Big(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ C_{t-1} + b_i\Big) \tag{8.3}$$

where $W_{xc}$ and $W_{hc}$ are the learnable weight matrices applied over the current input data $x_t$ and previous hidden state $h_{t-1}$, respectively; $W_{xf}$, $W_{hf}$ and $W_{cf}$ are the weights applied in the feature vector, previous hidden state and previous cell state in the forget gate, respectively; $W_{xi}$, $W_{hi}$ and $W_{ci}$ are the weights applied in the feature vector, previous hidden state and previous cell state in the input gate, respectively.

Further, an output gate $o_t$ decides whether $C_t$ will be propagated to the final state $h_t$ according to the following expressions:

$$o_t = \sigma\Big(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \circ C_{t-1} + b_o\Big) \tag{8.4}$$

$$h_t = o_t \circ tanh\Big(C_t\Big) \tag{8.5}$$

where $W_{xo}$, $W_{ho}$ and $W_{co}$ are the learnable weight matrices applied over the input data, previous hidden state and cell state in the output gate, respectively. In Equations 8.1 to 8.5, symbols $*$ and $\circ$ denote the convolution operator and Hadamard product, respectively.

In the proposed architecture, the convolutional LSTM units are arranged in an encoder-decoder configuration in order to handle the temporal information from the activations delivered by the prior block, i.e. the last convolutional layer in the spatial encoder. The temporal encoder maps the feature sequence to a vector of fixed dimension based on all inputs of the sequence and therefore takes advantage of its temporal structure. The resulting cell and hidden states contain the learned motion from the sequence. This representation is fed to initialize the internal states of the temporal decoder where another LSTM decodes the target sequence from the vector.

The spatial decoder consists in a generative pathway. It is composed by three up-convolution layers to upsample the features back to the original dimension. Before every up-convolution, features coming from the spatial encoder are concatenated to the decoder through skip connections. Using this scheme, the up-sampling path is provided with information captured in early layers that may be useful for the reconstruction. Five convolutional layers, with same configuration as in the encoder (in terms of kernel size and stride), reduce progressively the feature maps. In this case, its final size can represent a bidimensional space, namely superior-inferior and anterior-posterior motion.

### 8.2.3 Spatial transformation module

Spatial Transformer Networks (STN) were proposed by [161] to provide Convolutional Neural Networks with explicit spatial transformation capabilities. This module performs spatial warping on images or feature maps by producing a transformation with learnable parameters. It also contains a differentiable grid sampling function allowing for backpropagation within an end-to-end learning framework. Inspired by this, we introduced a cascade of spatial transformation functions at the end of the model to sequentially yield the predicted image sequence. Unlike the original STN, where the transformation is intrinsically learned by internal neural networks, our model infers the transformation from the last convolutional layer of the decoder which provides the sequence of predicted flows. Starting by the last image in the input sequence $I_n^i$, the first spatial transformer warps it with $\phi_n$ to produce $I_{n+1}^o$. Subsequently, the resulting image $I_{n+1}^o$ is warped with $\phi_{n+1}$ by another independent spatial transformer yielding $I_{n+2}^o$ and so on, for all the predicted time steps. An alternative approach is to perform one-shot warping on the last input image $I_n^i$ with the algebraic sum of the deformations over $T$ times to get the output image:

$$I_{n+T+1}^o = I_n^i \circ \sum_{\tau=0}^{T} \phi_{n+\tau}. \tag{8.6}$$

This can be beneficial in the presence of noise, e.g. US images, where speckle noise can deteriorate image quality during iterative warping.

Each spatial transformation takes the form of a smooth dense displacement field $u$. Following the common notation used in medical image registration, we represent the transformation at every pixel position $\boldsymbol{p} \in \mathbb{R}^{H\times W}$ as the summation of an identity transformation with the displacement field $u$, or $\phi(\boldsymbol{p}) = \boldsymbol{p} + u(\boldsymbol{p})$. For each $\boldsymbol{p}$ in the source image $I_n$ at time $n$, we compute a subpixel resolution location $\boldsymbol{p\prime}$ in the warped image $I_{n+1} = I_n \circ \phi_n$. In order to obtain new locations, we interpolate linearly the values at the neighboring pixels following [280]:

$$I_{n+1}\left(\boldsymbol{p\prime}\right)=I_n\circ\phi\left(\boldsymbol{p}\right)=\sum_{\boldsymbol{q}\in\mathcal{Z}\left(\boldsymbol{p\prime}\right)}I_n\left(\boldsymbol{q}\right)\prod_{d\in\mathbb{R}^2}\left(1-\left|\boldsymbol{p\prime}_d-\boldsymbol{q}_d\right|\right), \tag{8.7}$$

where $\mathcal{Z}\left(\boldsymbol{p\prime}\right)$ are the pixel neighbors of $\boldsymbol{p\prime}$, and $d$ iterates over a bidimensional space.

### 8.2.4 Similarity-based loss function

Similarly to the image registration domain, we propose a loss function to the neural network which measures the fit between the predicted and target images. Let $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ denote the target and predicted image sequences, respectively, and let $\mathbf{\Phi}$ be the sequence of predicted dense displacement fields. Our optimization problem can be written as:

$$\hat{\mathbf{\Phi}}=\arg\min_{\mathbf{\Phi}}\mathcal{L}_{total}\left(\mathbf{Y},\hat{\mathbf{Y}},\mathbf{\Phi}\right) \tag{8.8}$$

where

$$\mathcal{L}_{total}=\frac{1}{T}\sum_{k=1}^{T}\mathcal{L}_{sim}\left(I_k^t,I_k^p\right)+\lambda\mathcal{L}_{smooth}\left(\phi_k\right) \tag{8.9}$$

Here, $\mathcal{L}_{sim}\left(\cdot,\cdot\right)$ is a modular metric that measures the similarity between the target and predicted images, $I_k^t$ and $I_k^p$, respectively. $\mathcal{L}_{smooth}\left(\cdot\right)$ acts as a regularization term weighted by $\lambda$ on the deformation field $\phi_k$. The proposed framework is agnostic to the loss function as long as it be differentiable. Since the sequential images always come from the same modality, in this work the metrics used for $\mathcal{L}_{sim}$ include mean squared pixel difference and cross-correlation. Normalized Cross Correlation (NCC) is able to cope with intensity variations often found across scans and datasets. We minimize the negative local cross-correlation between predicted and target sequences by using exclusively convolutional operations similarly as proposed by [42]. Instead of computing local means over volume patches, we look at $n^2$ vicinities around each pixel with $n=9$:

$$NCC\left(I_k^t,I_k^p\right)=\sum_{\boldsymbol{p}\in\mathbb{R}^2}\frac{\left(\sum_{p_i}\left(I_k^t\left(\boldsymbol{p_i}\right)-\hat{I}_k^t\left(\boldsymbol{p}\right)\right)\left(I_k^p\left(\boldsymbol{p_i}\right)-\hat{I}_k^p\left(\boldsymbol{p}\right)\right)\right)^2}{\left(\sum_{p_i}\left(I_k^t\left(\boldsymbol{p_i}\right)-\hat{I}_k^t\left(\boldsymbol{p}\right)\right)\right)\left(\sum_{p_i}\left(I_k^p\left(\boldsymbol{p_i}\right)-\hat{I}_k^p\left(\boldsymbol{p}\right)\right)\right)} \tag{8.10}$$

where $\hat{I}_k^t$ and $\hat{I}_k^p$ are images in which local mean intensities have been subtracted.

Finally, we introduce a regularizer term $\mathcal{L}_{smooth}\left(\cdot\right)$ in the cost function to achieve smooth deformation fields and avoid discontinuities. It is modeled as a linear operator on spatial gradients of $\phi$:

$$\mathcal{L}_{smooth}\left(\phi\right) = \sum_{\boldsymbol{p}\in\mathbb{R}^2} \left\|\nabla\phi\left(\boldsymbol{p}\right)\right\|^2 \tag{8.11}$$

The spatial gradients of the deformation vectors $\nabla\phi$ are calculated from the differences between neighboring locations of $\boldsymbol{p}$. The regularizer term $\mathcal{L}_{smooth}\left(\cdot\right)$ was weighted with $\lambda$ equal to 0.01 and 1.0 when used with the MSE and NCC similarity metrics, respectively.

### 8.2.5 Training protocol

Datasets were divided in different subsets for training, validation and testing purposes. Hence during testing the model performs predictions from unseen anatomies. The proportions used for each dataset are specified in Section 8.3.1. Mean centering and standard deviation normalization was applied to each input image. This decreases the variability of the input data, thus improving the training stability. Adam optimizer was used with an initial learning rate of $10^{-3}$. This learning rate was reduced by a factor of 2 after 5 epochs without improvements in the validation set accuracy. The allowed minimum learning rate was $10^{-10}$. Finally, early stopping was used to prevent overfitting. We implemented our model in the Keras framework [281] using Tensorflow backend [282].

## 8.3 Experiments and results

In this section, we present the experimental results of the proposed pipeline for medical sequence prediction. We first provide details on the different dataset used to validate the model. Furthermore, we describe the metrics used for evaluation purposes as well as the implementation of comparative methods. We then present quantitative results on the comparison of our method with related approaches through different imaging modalities. It is important to highlight that in our comparisons, we are interested in assessing the predictive capacity of the model. Consequently, we selected state-of-the-art approaches performing the same in-plane motion prediction task. Finally, we show some qualitative results in each case for three modalities (MRI, US, CT) and the computational times with the hardware used.

### 8.3.1 Datasets

**Magnetic resonance images** This dataset consists of liver sagittal scans without contrast agent from 12 healthy subjects. This study was approved by the institutional review board and written consent was obtained from all the volunteers. Volunteers were instructed to breathe normally during the entire acquisition. Cine-phase images were acquired on a Siemens

Skyra 3T scanner using a 2D T2-weighted true FISP sequence with pixel matrix of 176 × 176, pixel spacing of 1.7 × 1.7 mm$^2$ and a slice thickness of 3 mm. The temporal resolution was 320 ms. In order to avoid mixing different movement patterns (breathing induced with cardiac beating) the slices covered only the right hemi-diaphragm leaving out the cardiac cavity. Each anatomical position comprises 50 dynamics. We followed a leave-one-out validation scheme on a subject level. Thus, the model was tested on all the slice positions belonging to an unseen subject. As a pre-processing step, images were cropped to 112 × 112 to focus on the liver and remove organs in the bottom part such as stomach, kidneys, pancreas, intestines, etc. Figure 8.4 shows the liver motion range measured from vessels positions for each subject during the free breathing sequence. These values are in concordance with the motion ranges reported in [22]. In addition, we observed that the MRI dataset contains different motion amplitudes, ranging from shallow up to deep motion patterns, which increases the inter-subject variability.

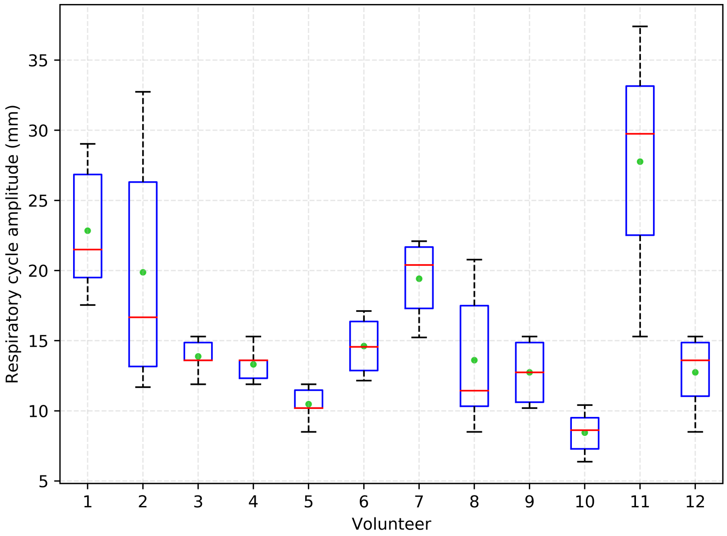


Figure 8.4 Observed motion in the MRI dataset from vessel tracking. Box plot shows the minimum value, the first quartile, the median, the third quartile and the 95$^{th}$ percentile of the values. Mean values are marked by green points.

**Ultrasound images** The CLUST dataset [283] was originally created for 2D and 3D liver vessel tracking challenges. It contains 63 free-breathing sequences acquired at various centers with different scanners. Image sizes, pixel and temporal resolutions, ranging between 262 × 313 and 475 × 687, 0.27 mm and 0.77 mm and 32 ms (31 Hz) and 90 ms (11 Hz). Manual annotations of anatomical landmarks were provided. Images were preprocessed before the training, validation and testing. Consequently, pixel resolutions were normalized to 0.5 × 0.5 mm$^2$. This value represents the mean value of the original resolutions range. Zero-padding was applied in order to obtain a consistent image size ($480 \times 640$ pixels) through all the data. Similarly, image rates were normalized to 200 ms (5 Hz). From the total of 63 sequences, 39 were used for training, 8 for validation and 16 for testing which represents, approximately, 62%, 13% and 25% of the total, respectively.

**Computed tomography images** The third and final dataset to be used was 4D thoracic CT dataset [284, 285] which is publicly available at `http://www.dir-lab.com`. It contains thoracic CT images from 10 patients which were acquired as part of the treatment planning process for the treatment of thoracic malignancies. Anatomical landmarks were manually identified by an expert in thoracic imaging, with additional registration performed by multiple observers. Similarly as the MRI dataset, we used fifty slices covering the right hemi-diaphragm with 10 dynamic each. Pixel spacing was normalized to $1.0 \times 1.0$ mm$^2$. The temporal resolution is approximately 400 ms. Similarly as the previous datasets, image acquisition was carried-out during free-breathing.

### 8.3.2 Evaluation metrics

We evaluate the performance of the proposed method both quantitatively and qualitatively. To perform the quantitative evaluation, we analyzed the local and global behavior using landmark errors and image similarity metrics, respectively. The local behavior was assessed by measuring the distance between landmarks in the ground-truth and predicted images. To this end, manual annotations of blood vessel positions and bifurcations provided by experts were used for MRI, US and CT datasets. These landmarks were located in the predicted images using the Lucas-Kanade optical flow algorithm implemented in OpenCV library [286] to find the estimated positions. The prediction error for each landmark $n$ in a temporal image $t$ was measured as the Euclidean distance $E_{n,t+\Delta} = \|g_{n,t+\Delta} - p_{n,t+\Delta}\|$ between target $g$ and predicted $p$ positions. For certain experiments we used the Normalized Cross Correlation to capture the global spatial coherence between ground truth and predicted images.

### 8.3.3 Implementation of comparative methods

Both comparative methods require deformable image registration as part of their processes: the motion matrix construction for PCA, and the label creation for the deep approach. Therefore, pairwise deformable registration was performed using the publicly available tool NiftyReg [201] in order to quantify the motion between consecutive temporal images. Dense displacement fields were obtained by using a cubic B-Splines transformation model in a pyramidal approach. A source image was iteratively deformed while optimizing an objective function based on the Normalized Mutual Information with 64 bins and a penalty term based on the bending-energy [30].

**Statistical modeling:** Deformable registration was accomplished following the previously mentioned configuration after the rigid alignment between a reference phase (exhale in this case) and the rest of phases. The resulting displacement fields were utilized to construct PCA motion models for comparison purposes. Specifically, individual models using two principal components were created for each anatomical position and its temporal images. The explained variance with two components was [0.96, 0.03].

Figure 8.5 illustrates the methodology used to produce the spatio-temporal predictions. Recent state-of-the-art works have used surrogates to drive the models [254, 260]. One dimensional navigators extracted from a tracking method presented by [80] were used as surrogates. They were linked to the PCA coefficients through linear regression. Adaptive linear filtering was used to extrapolate respiratory signal values in time. According to [260] this approach yielded the best performance among several compared temporal prediction methods. Updated PCA coefficients were derived from the known relationship between model and surrogate. Finally, these coefficients were used to feed the model and to produce the spatio-temporal predictions.

**Classification-based approach (uniform grid)**: We also implemented the classification-based motion prediction approach proposed by [9], where a method is used to uniformly quantize the range of values for each vectorial component. In our case, these ranges represent the anterior posterior (AP) and superior inferior (SI) motion planes. Using this uniform scheme, the values covered by each bin were determined by dividing the whole range (max-min) by the number of bins. Then, different classes were assigned to each possible combination between the bins of each motion plane. The number of bins was equal to 5, yielding 25 classes ($5 \times 5$). Weighted cross-entropy was used as loss function as the original paper [9].

**Classification-based approach (adaptive grid):** Figure 8.6 shows the probability distribution of the motion vectors obtained from deformable registration in the MRI dataset. Its approximately-Gaussian shape is a key characteristic that must be taken into account during the space discretization. Therefore, instead of dividing uniformly the range with the number of bins, we follow an alternative scheme which we refer as adaptive grid (AG). It consists of selecting the bins near to the mean, standard deviation, minimum and maximum distribution values. Using this scheme, we effectively represent the motion distribution observed in the dataset.

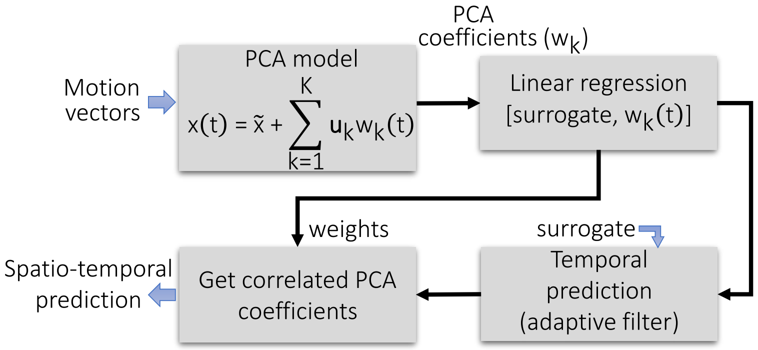

PCA coefficients ($w_k$)
Motion vectors
PCA model
$x(t) = \tilde{x} + \sum_{k=1}^{K} \mathbf{u}_k w_k(t)$
Linear regression [surrogate, $w_k(t)$]
weights
surrogate
Spatio-temporal prediction
Get correlated PCA coefficients
Temporal prediction (adaptive filter)

Figure 8.5 Implemented PCA pipeline for in-plane spatio-temporal prediction.

Table 8.1 Vessel tracking error position (in mm) for each predicted time in the MRI dataset. Values are median (interquartile range).

| Model | $t_p$=1 | $t_p$=2 | $t_p$=3 | $t_p$=4 | $t_p$=5 |
|---|---|---|---|---|---|
| | (320 ms) | (640 ms) | (960 ms) | (1280 ms) | (1600 ms) |
| Classification | 1.63 (2.29) | 2.32 (2.58) | 2.96 (2.88) | 3.23 (2.65) | 3.55 (2.52) |
| Classif. (AG) | 1.55 (1.45) | 2.33 (2.10) | 2.77 (2.64) | 3.16 (2.63) | 3.20 (2.82) |
| PCA | 1.36 (2.73) | 1.85 (2.98) | 2.37 (2.88) | 2.72 (2.67) | 3.01 (2.49) |
| ED-ST(ncc) | 0.54 (0.66) | 0.74 (0.98) | 1.03 (1.26) | 1.17 (1.42) | 1.30 (1.66) |
| MSED-ST(ncc) | 0.43 (0.54) | 0.72 (0.91) | 0.88 (1.22) | 1.01 (1.36) | 1.21 (1.57) |
| ED-ST(mse) | 0.56 (0.65) | 0.77 (0.96) | 0.94 (1.15) | 1.00 (1.15) | 1.28 (1.61) |
| **MSED-ST(mse)** | **0.45 (0.55)** | **0.57 (0.75)** | **0.80 (0.99)** | **0.88 (1.25)** | **0.77 (1.36)** |

### 8.3.4 Results

We now present a comparison between multiples variants of our pipeline and state-of-the-art approaches for spatio-temporal prediction, which were described above, namely: PCA and

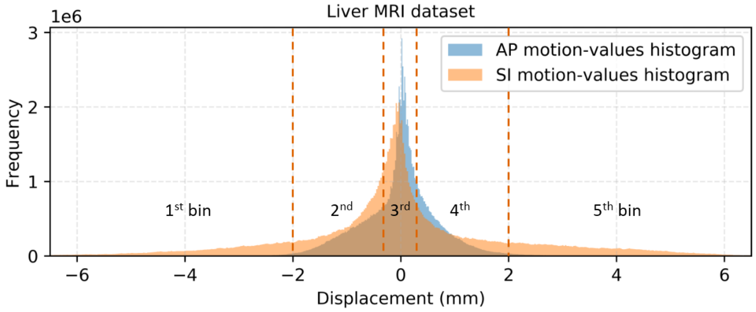


Figure 8.6 Histograms of the superior-inferior and anterior-posterior displacements in the liver MRI dataset. Dashed lines delimit the bins. For clarity only the division of the SI component is represented.

classification-based models. As explained in Section 8.3.3, our classification-based implementation follows the method introduced by [9]. Moreover, we explore the benefits of the proposed adaptive motion encoding.

Our experiments are aimed at investigating three aspects regarding the time extrapolation and the proposed multi-scale feature extraction model: (1) the importance of the multi-scale architecture design, (2) the effect of increasing the number of predicted frames on the overall accuracy of the model, and (3) the effect of varying the number of prior frames on the overall accuracy of the model. To address these effects, we ran paired experiments with and without the multi-scale architecture. Therefore, we investigate the performance of: (1) stacked convolution and pooling layers for feature extraction and (2) multi-scale residual feature extraction, hereinafter referred as "ED-ST" and "MSED-ST", respectively. We also present results on the use of different loss functions. Predictions were extrapolated at $\{1, 2, 3, 4, 5\}$ time points given the same number of prior time steps at the input (5 images for all the cases). Statistical significance was calculated by applying a Wilcoxon signed-rank test using the function implemented in the Python Scipy library. $P < 0.01$ was considered to indicate a statistically significant difference.

Tables 8.1, 8.2 and 8.3 list the median landmark tracking errors and interquartile ranges among the compared methods for MR, US and CT images, respectively. These tables present results on multiple predicted times. In these cases (number of predicted frames $> 1$), the reported value is the average of the individual errors at each predicted image. Seemingly, the reported distributions exhibit a trend in their values: error values increase as we generate more frames. It is natural that, based on the same information, the error increases when extrapolating more time points. In all cases the proposed approach provides the top accuracy for the landmark tracking. This behavior is consistent through all the time-resolved experiments and imaging modalities. As reported in the tables, landmark localization errors

Table 8.2 Landmark localization error (in mm) for each predicted time in the US dataset. Values are median (interquartile range).

| Model | $t_p$=1 | $t_p$=2 | $t_p$=3 | $t_p$=4 | $t_p$=5 |
|---|---|---|---|---|---|
| | (400 ms) | (800 ms) | (1200 ms) | (1600 ms) | (2000 ms) |
| Classification | 0.96 (1.35) | 1.33 (1.76) | 1.94 (1.77) | 2.22 (1.85) | 2.34 (1.66) |
| Classif. (AG) | 0.71 (0.92) | 1.01 (1.38) | 1.33 (1.57) | 1.43 (1.58) | 1.72 (1.64) |
| ED-ST(ncc) | 0.55 (0.83) | 0.81 (1.13) | 1.18 (1.41) | 1.18 (1.27) | 1.48 (1.52) |
| MSED-ST(ncc) | 0.51 (0.76) | 0.76 (1.08) | 1.03 (1.28) | 1.17 (1.24) | 1.25 (1.28) |
| ED-ST(mse) | 0.49 (0.82) | 0.75 (1.20) | 1.03 (1.23) | 1.24 (1.30) | 1.44 (1.34) |
| **MSED-ST(mse)** | **0.45 (0.74)** | **0.74 (1.16)** | **0.98 (1.20)** | **1.18 (1.28)** | **1.28 (1.31)** |

Table 8.3 Landmark localization error (in mm) for each predicted time in the CT dataset. Values are median (interquartile range).

| Model | $t_p$=1 | $t_p$=2 | $t_p$=3 | $t_p$=4 | $t_p$=5 |
|---|---|---|---|---|---|
| | (400 ms) | (800 ms) | (1200 ms) | (1600 ms) | (2000 ms) |
| Classification | 1.03 (1.03) | 1.55 (1.63) | 1.98 (2.23) | 2.33 (3.00) | 2.75 (3.49) |
| Classif. (AG) | 0.72 (1.04) | 1.22 (1.53) | 1.59 (2.02) | 1.85 (2.15) | 2.14 (2.54) |
| ED-ST(ncc) | 0.33 (0.57) | 0.45 (0.64) | 0.52 (0.69) | 0.54 (0.69) | 0.57 (0.73) |
| MSED-ST(ncc) | 0.25 (0.52) | 0.38 (0.53) | 0.42 (0.55) | 0.44 (0.58) | 0.45 (0.55) |
| ED-ST(mse) | 0.32 (0.61) | 0.42 (0.56) | 0.47 (0.52) | 0.46 (0.52) | 0.49 (0.53) |
| **MSED-ST(mse)** | **0.28 (0.58)** | **0.37 (0.56)** | **0.41 (0.52)** | **0.42 (0.47)** | **0.42 (0.49)** |

ranges between 0.45 (0.55) and 0.77 (1.36), 0.45 (0.74) and 1.28 (1.31) and 0.28 (0.58) and 0.42 (0.49) for one and five predicted time steps in MRI, US and CT datasets, respectively.

When comparing results for overall landmark errors in one predicted time with the model which did not integrate multi-scale feature extraction, the accuracy improved by a statistically significant margin of 0.11 mm, 0.04 mm and 0.04 mm ($p<0.01$), of 1.1 mm, 0.26 mm and 0.44 mm ($p<0.01$) to a classification-based approach and of 0.91 mm ($p<0.01$) to a statistical model, for MRI, US and CT dataset, respectively. These results seem to confirm two-fold outcomes: first, using a multi-scale approach does help to converge towards an optimum minimum, and second, minimizing image differences through a regression model does improve the current state-of-the-art for the sequential prediction task.

Reported results reveal that PCA performed better than classification-based models. As hypothesized, the adaptive gridding achieved better performance among the compared classification methods in comparison with the approach presented by [9]. As shown in Figure 8.6, there is a high density in the central part of the distribution. In consequence, the adaptive grid will recover approximate values nearest to the actual values improving the results.

We performed additional experiments on MRI slices belonging to the left liver lobe to test the model capacity to learn composite motion (see Figure 8.7). Figure 8.8 shows the NCC values obtained with the proposed model (MSE-based optimization) following a leave one out scheme on the 12 subjects. Reported results reveal a performance slightly lower than those obtained in slices where most of the movement was breathing-induced, but remain quite accurate. This showcases the complexity of predicting composite motion from various sources. In this case, the images capture respiratory, cardiac and even peristaltic motion. Nevertheless, mean NCC values higher than 0.85 for the longest extrapolation (5 predicted times) seem to confirm that the model extracts useful spatiotemporal latent features which allows to learn and predict the complex underlying dynamics.

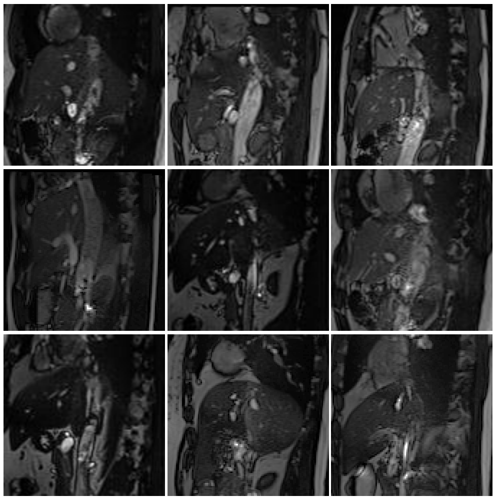


Figure 8.7 MRI slices belonging to the left liver lobe from several subjects.

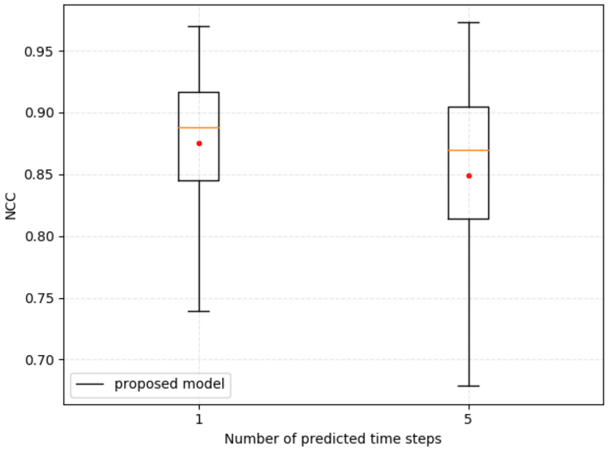


Figure 8.8 NCC between target and predicted images of the left liver lobe using the proposed model (MSE-based optimization) for 1 and 5 timesteps.

Figure 8.9 displays the vessel trajectory through the target and predicted temporal MR images. Our multi-scale regression-based model trained with MSE as loss function showed a near perfect alignment with the target trajectory. Therefore, this confirms that our procedure is a clear improvement on current methods.

Figure 8.10 presents NCC values for all different predicted times $\{1, 2, 3, 4, 5\}$ when varying the number of input images $\{2, 3, 4, 5\}$ given to the proposed model using the MRI dataset. The case with one input image was avoided to maintain the sequence to sequence approach. Results were obtained following a leave one out scheme on the 12 cases and show an increased performance when more input images are provided.

Finally, we present qualitative results on the adopted datasets. Figure 8.11 shows the difference maps between the ground-truth and predicted images obtained with the classification, PCA and the proposed models. Images predicted by our framework match target images showing pixel-wise errors near to zero, particularly in regions with vessels. Multiple time predictions on this dataset are shown in Figure 8.12. Images were selected at challenging end phases such as exhale. The yellow boxes indicate the most noticeable inconsistencies among the compared images. Intermittent blood vessels that appear suddenly as a result of the out-of-plane movement are not captured in the predictions. This is mainly due to two factors: (1) the appearance content depends on the last image seen in the input sequence and (2) the model is not generative by nature but discriminative.

Figure 8.13 shows two inference cases from the ultrasound dataset: (a) next frame prediction

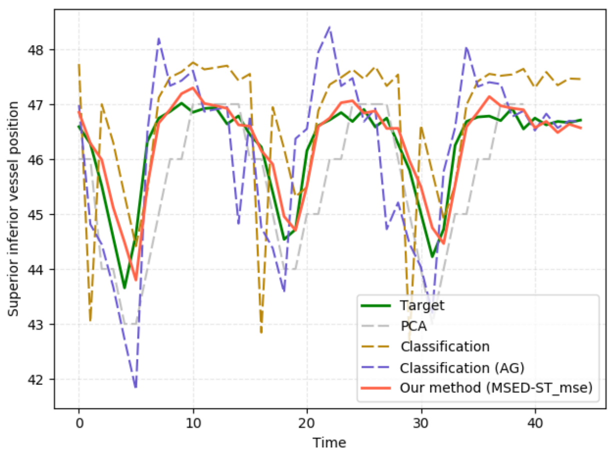


Figure 8.9 Vessel trajectory from images predicted with different approaches for one case in the MRI dataset.

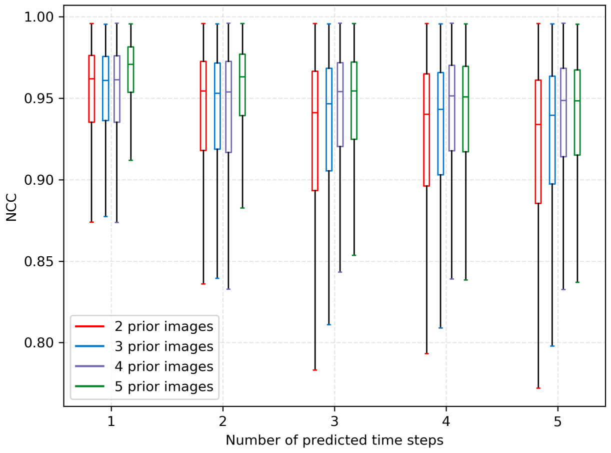


Figure 8.10 NCC between target and predicted MRI images when varying the number of prior images provided to the proposed model (MSE-based optimization).

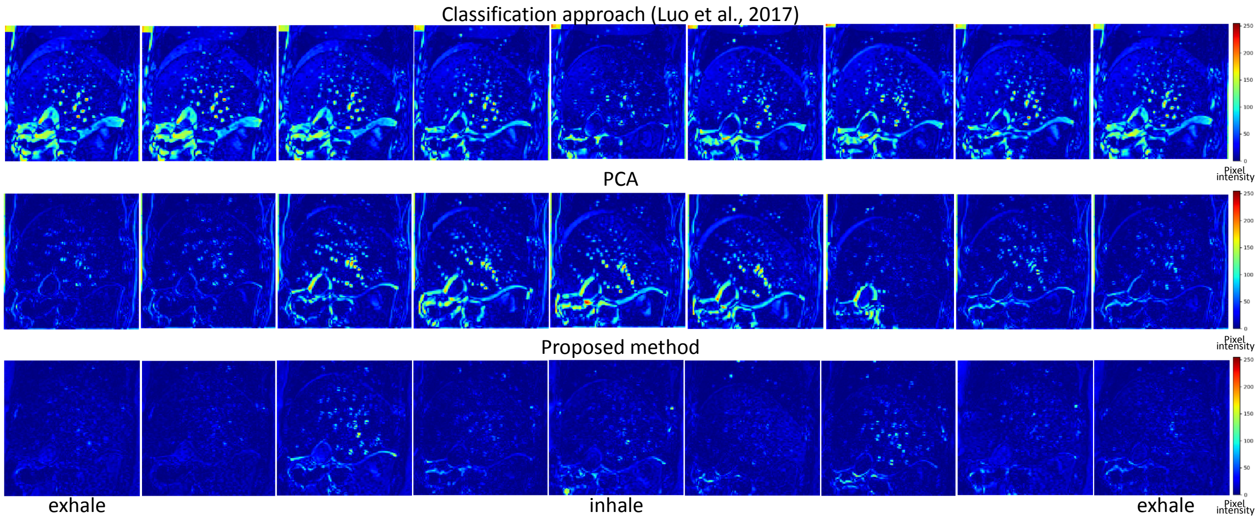


Figure 8.11 Difference maps between ground-truth and predicted MR images from different methods through a complete respiratory cycle. Color bars indicate the pixel intensities of a 8-bit grayscale image which range from 0 (blue) up to 255 (red).

and (b) five frames prediction. In Figure 8.13 (a) we observe that slight misalignment in the diaphragm edge occurs in the predicted frame, which correspond to inhale phase. This observation, also evident in Figure 8.11, suggests that errors are more prone to happen over that specific phase. This is probably due to large displacements during air intake within the lungs. In the case of more than one predicted time, a single model inference is shown. The longer it is extrapolated, the more image becomes blurred. Presumably, this is due to the

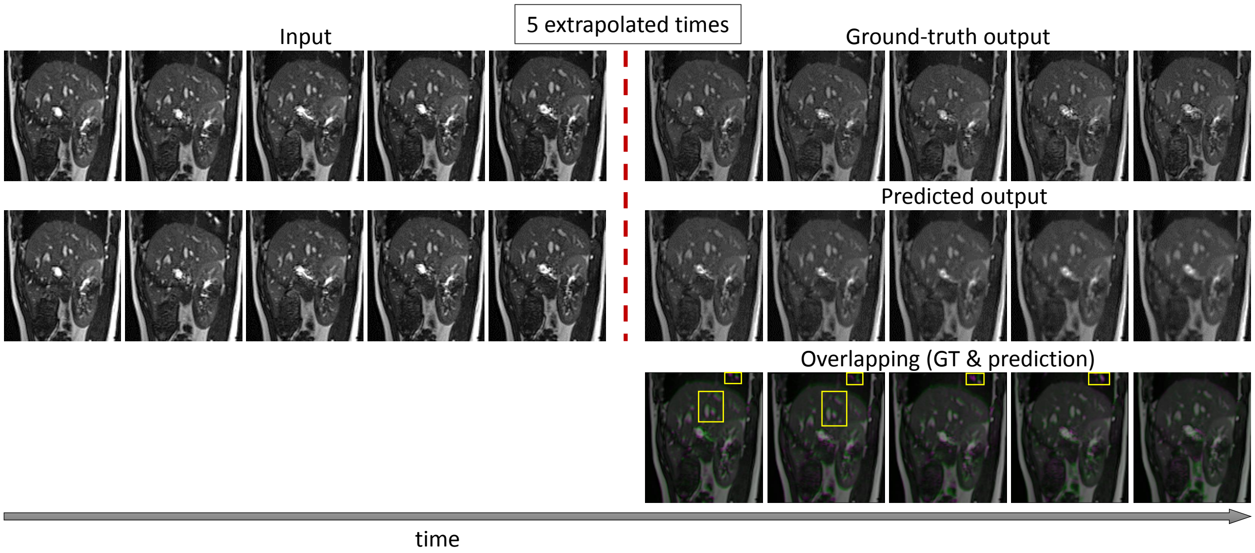


Figure 8.12 Extrapolation up to five time steps from an input sequence belonging to the test set in the MRI dataset. Top row: Target images, middle row: predicted images by the proposed model, bottom row: overlapping between target and predicted images. Magenta and green pixels belong to target and predicted images, respectively. Yellow boxes indicated some vessel inconsistencies due to out-of-plane motion.

effect of the successive transformations and the speckle noise contained inside the US images. Nonetheless, even though the last images are not as sharp, landmarks are easily identifiable. The use of equation 8.6 for warping is an alternative that may improve the quality as it always uses the last image in the sequences as source image. In our experiments we did not exploit such alternative as the generated images maintained an acceptable quality. In the illustrated examples, mismatches becomes evident at the liver edge, as shown in the yellow boxes.

It should be noted that different ultrasound probes from different commercial scanners were used and were placed in different positions through the dataset. For example, the images showed in Figure 8.13 (a) and (b) were acquired with Siemens and Clarity equipment. Moreover, as can be observed, in one case (Figure 8.13 (a)) the liver edge appears in the opposite orientation to the other Figure 8.13 (b). Consequently, there is a higher variability between scans in terms of organ orientation and appearance. This indicates the generalization capability of the model. Finally, Figure 8.14 presents qualitative results on multiple time predictions on CT dataset. In comparison with other imaging modalities, the predictions seems closer to target images. Nevertheless, this dataset captures only an average movement. This means that inter-cycle variability, which is a challenging factor, is not taken into account.

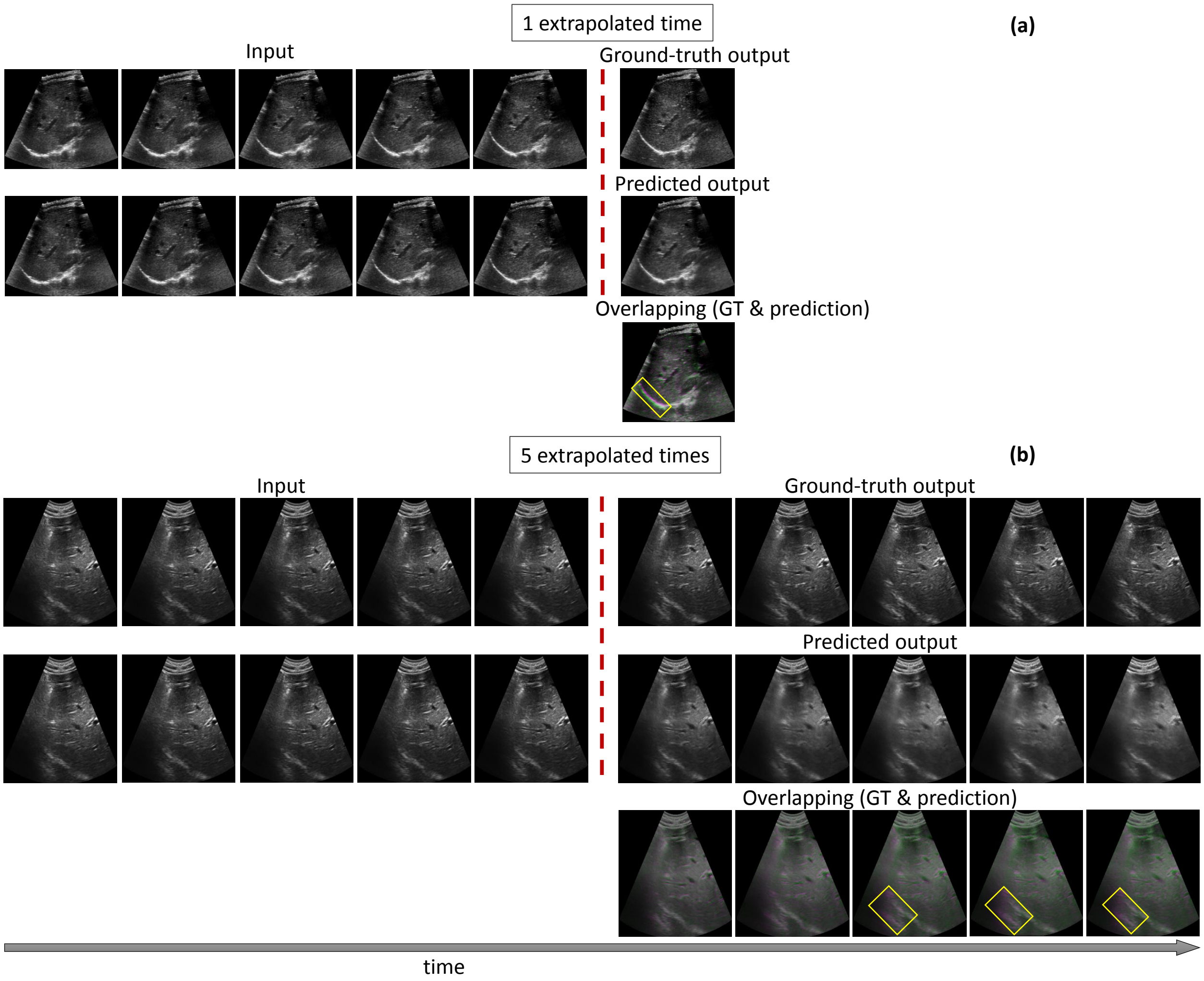


Figure 8.13 Extrapolations obtained from a single input across different views and scanners in the ultrasound dataset (sequence ID, scanner): (a) one time (ETH–04–02, Siemens Antares) (b) five times (ICR–01, Elekta Clarity) are showed in the upper and bottom parts, respectively. In both cases, the top row contains the input sequence and the expected target images; the middle row contains the input sequence and the predicted images by the proposed model; last, the bottom row shows the overlapping between target and predicted images. Magenta and green pixels belong to target and predicted images, respectively. Yellow boxes highlight some inconsistencies due to out-of-plane motion.

### 8.3.5 Computational times

The training of the proposed model took approximately 2 hours in the largest dataset (MRI), leaving out the test case. We report the following times for the total number of images in a single network inference during testing as a result of averaging 10 individual measurements: 35 ms, 35 ms, 38 ms, 38 ms and 39 ms for 1, 2, 3, 4 and 5 predicted time steps, respectively.

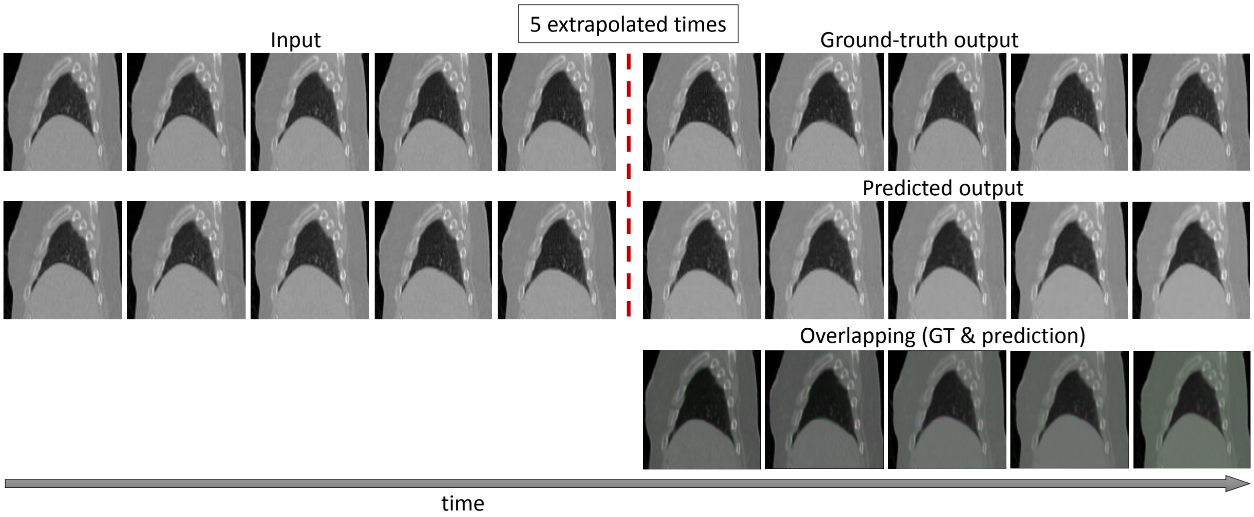


Figure 8.14 Extrapolation up to five time steps from an input sequence belonging to the test set in the CT dataset. The top row contains the input sequence and the expected target images; the middle row contains the input sequence and the predicted images by the proposed model; last, the bottom row shows the overlapping between target and predicted images. Magenta and green pixels belong to target and predicted images, respectively.

All measurements refer to a Python implementation running on a machine equipped with a 3.50 GHz processor, 64 GB RAM and a GPU NVIDIA GeForce GTX TITAN X.

## 8.4 Discussion and conclusion

This paper presents the first multi-scale recurrent encoder-decoder framework able to generate future image sequences based on iterative spatial transformations from free-breathing scans. The reported results on three different modalities (MRI, ultrasound, CT) demonstrate that the proposed method outperforms state of the art methods for in-plane motion prediction in terms of spatial and temporal quality. Compared to the statistical modeling methodologies proposed in the literature, our method does not need any pre-processing step such as segmentation or displacement field extraction.

In several previous methods, organ segmentation is performed prior to image registration to isolate the organ motion. Generally, it is verified manually to avoid uncertainties introduced by automatic contouring methods. On the other hand, nowadays there are powerful tools for deformable image registration which have demonstrated excellent performance. However, this process can be time-consuming, especially in large datasets. In our model, the alignment between consecutive images is performed implicitly during training. Moreover, once trained, the model weights remain optimized to help generalize for unseen images. In other approaches, the registration is an independent and mandatory step prior to inference, whether to construct the motion matrix (PCA) or to create target labels (classification approach). From our point of view, an implicit registration compared to prior-step registration presents the advantage where inference can be done in near real-time. For example, for PCA modeling, registration on each new patient dataset must be done before. Instead of that, the proposed model learns global alignment parameters allowing one-shot prediction. While classical registration approaches still outperform deep learning techniques in several medical imaging applications, we observed in registration cases where cyclical organ motion is apparent on image sequences of free-breathing images, the deep learning approaches tend to better capture this phenomenon by learning latent features related to internal motion. On the other hand, in the case of population-based models, finding inter-subject correspondences between anatomical landmarks is an additional step in the pipeline towards the statistical model creation, which can lead to inaccuracies and poor repeatability.

Our method also has significant benefits over the classification-based model. Such approaches introduce an inherent error during the motion encoding for the label creation. That error depends directly on the number of bins selected to cluster the deformation vectors. On the other hand, potential misclassification may lead to unrealistic and ambiguous motion. Similarly to statistical models, image registration is a mandatory step during label creation. It might be time-consuming, depending on the available computational resources. Additionally, during deployment it is necessary to recover the actual deformation values from the predicted motion classes. In contrast to the aforementioned, our proposed method is trainable end-to-end and fully unsupervised. No previous segmentation, registration nor intermediate motion encoding/decoding are needed. Furthermore, it presents other attractive and practical characteristics. It can be trained with images from different subjects, anatomies, scanners and views. As demonstrated in this study, our framework can learn the motion patterns regardless the imaging modality which makes it very valuable for clinical applications. Thanks to the feature extraction stage, the network learns in the domain of pixel intensities and implicitly regress the transformations. It simultaneously learns to register the images and subsequently to extrapolate the deformation in time. The experiments showed the model performance with increasing number of predicted times, ranging from one to five. Also, the inference time was reported for each case. It is important to note that the optimal predicted number of images will depend heavily on the clinical application. While the top performance was obtained when predicting the next frame, some applications may require additional extrapolations. Nonetheless, the proposed model showed a suitable behavior in all the cases.

Finally, some limitations remain which will need to be addressed in future studies. The first is related to the presence of large deformations. However, we believe it is unlikely to happen as we are modeling sequential deformations instead of deformation from a particular phase or state. The second is the inability to cope with out-of-plane motion. Generative models might be a solution and should be investigated in future work. Also, in order to solve this limitation, the extension to 3D prediction or modeling might be considered. Last, but not least, is the necessity of covering exhaustively the immense anatomical and structural variability that may be seen during the clinical routine before its actual introduction. Further studies should be concentrated towards optimization for real-time application.

**Acknowledgments** This work was partly funded by an NSERC collaborative research and development project (CRDPJ-517413-17) and by the Canada First Research Excellence Fund through the TransMedTech Institute.

# CHAPTER 9 ATTENTION-BASED TEMPORAL PREDICTION AND TRACKING

## 9.1 Introduction

In medical imaging, the analysis of temporal image sequences allows to examine physiological processes within the human body. Organs that undergone certain dynamic, such as the heart, liver or lungs, are particularly interesting to study whether for diagnostic purposes or for treatment delivery. Furthermore, modelling the motion patterns allows to track anatomical structures, to compensate for motion, to do temporal super-resolution and motion simulation [269]. In the context of image-guided interventions, modelling the temporal behavior allows to compensate for the system latencies. Certainly, processes such as image acquisition, target localization, and subsequent beam modulation/tracking result in a significant cumulative system latency. This means that, in real-time treatments, by the time a gating decision has been made, the patient anatomy has already changed. Therefore, temporal forecasting is required to compensate for these latencies.

Modern image-guided radiotherapy (IGRT) systems enable the acquisition of cine slices at certain anatomical position, which provides in-plane information. These intra-operative images can be used as partial observations to derive volumetric information from previously created motion models. Such volumetric information is useful to estimate tumor position and to monitor organ at risks throughout the treatment [191].

Motion models relate partial observations, such as 2D slices, with high-dimensional motion measurements (e.g. 3D deformation fields). Generally, two types of motion modeling can be distinguished: local and global [119]. Local approaches use information surrounding the target to reconstruct exclusively the 2D/3D position of the tumor, while global approaches estimates the entire 2D/3D anatomy. Local modelling is commonly used for target tracking through time. Similarly to visual tracking, in the general computer vision domain, this process refers to matching instances in consecutive time frames. However, compared to natural images, medical sequences present additional challenges such as complex background, variable target sizes and appearances. Currently, Siamese networks constitute the state-of-the-art for target tracking [287]. Nevertheless, this type of models requires both source and target images as inputs, which is not possible in real-time slice-based intra-treatment acquisitions. Moreover, the feature representation in Siamese networks does not take into account contexts of other objects and it does not fully utilize complementary information, for instance, motion. We argue that both low-dimensional and high-dimensional motion modelling outcomes can be useful to achieve refined 3D+$t$ target trajectories from cine slices.

On the other hand, classical approaches for global deformation estimation rely on biomechanical modeling, image registration and statistical modeling, the latter being the most common. Recently, some authors have proposed modelling frameworks leveraging the benefits offered by deep neural networks [180,190,191,288]. Since motion compensation is an important requirement, these motion models generally comprise mechanisms for temporal forecasting. Moreover, with the increasing imaging capabilities in the clinical units, the forecasting is expected to be performed on the image domain. In the literature, a vast amount of approaches for temporal prediction rely on recurrent neural networks (RNN) [10,119,134,138,144,165,166]. Long-Short Term Memory and their variants have been applied to sequential data learning, including natural language, video, among others tasks [104]. Particularly for spatiotemporal modeling, using a serie of Convolutional Long-Short Term Memory (ConvLSTM) [112] units is a popular choice. Nevertheless, it has been shown that ConvLSTM presents a blind-spot problem since it fails to consider the entire spatiotemporal context from previous frames [166]. Recently, Transformers have made breakthrough in many natural language processing tasks, where they have shown excellent performance for representation learning. Hence, there have been an increasingly interest in adapting this attention-based model for computer vision problems. However, in the medical imaging field their benefits for temporal forecasting have been less explored. Similarly to natural language, respiratory motion presents a sequential dependency, which can be leveraged as additional knowledge to regress future values.

In this work, the main contributions are as follows:

- We propose a self-supervised model to predict future representation from an image sequence by learning queries within a Transformer architecture.
- We leverage future frames, available during model training, to compute a prior over their latent representations.
- We propose a model-based region tracking strategy, which enables ahead-of-time 3D target tracking from image surrogates.

## 9.2 Related works

### 9.2.1 Motion prediction

Video forecasting is an active research topic in general computer vision [113,132,133,135,138,144,146,151,159,160]. A common characteristic of most of the proposed models is that they often follow an encoder-decoder architecture. Furthermore, during the encoding process, some methods learn disentangled representations by decomposing the video in appearance and motion [8, 135–139]. This strategy seems to be effective in object-centric datasets (e.g. Moving MNIST, KTH), where the static background can be easily divided from the foreground. However, other videos containing crowded scenes or medical sequences, might be harder to directly disentangle. In terms of backbone deep learning architectures, predictive models generally rely on RNN [9, 101, 112, 119], although convolutional neural networks (CNN) [120, 122] and generative models [105, 136, 167], such as variational autoencoders (VAE) and generative adversarial networks (GAN), have also been employed. However, RNNs present some well-known drawbacks, which limit the temporal information modeling ability inherently. For instance, vanishing gradients remain an issue when they are applied for long-term dependency learning, as the recursive prediction may accumulate the errors. In addition, the sequential processing makes it unsuitable for parallel computation. Recently, attention-based mechanisms, such as the so-called Transformer [7], emerged as a promising solution for computer vision tasks [126–128]. A comprehensive survey about Transformers for computer vision can be found in [123].

On the other hand, multi-frame predictions still represent a challenge. In many SOTA approaches, the first few predicted frames are sharp while the visual quality quickly degrades when increasing the predictive horizon. Furthermore, pixel synthesis is challenging due to the high-dimensionality of the images [104]. Therefore, the blurriness is even more remarkable in models performing direct pixel regression [131, 135, 143, 150–155]. This constitutes an important limitation, particularly in the medical field. Alternatively, other approaches generate future frames by applying spatial transformations on the previous frames, like warping [119,159,160]. Nonetheless, it is important to highlight that this process is not exactly the same in natural images compared to medical images, where the motion observed in an image sequence is typically described by deformation fields between a source image and a fixed image. This presents an additional challenge, since the model must learn in the pixel intensity domain and implicitly regress the future transformations. In other words, it simultaneously learns to register the images and to extrapolate the deformations ahead-of-time [119].

From a probabilistic point-of-view, most of the works on motion prediction can be considered as deterministic [113, 132, 133, 142, 146, 151, 165], meaning that the model will always provide the same results for a given input. However, for multiple equally plausible outcomes, these models tend to average the result. In contrast, stochastic models usually generate sharper predictions [104]. For instance, Denton et al. [144] proposed to learn a probabilistic prior from future images recursively. Nonetheless, for some applications is desirable to obtain parallel outcomes rather than sequentially. In summary, information modeling for future frame prediction is still an open challenge due to the high-dimensionality inherent of real world videos. Predictions still suffer from lacking of high-frequency details and insufficient use of motion information [133].

### 9.2.2 Object detection and tracking

A vast amount of works have been dedicated to object tracking in diverse datasets using several paradigms [128, 289–293]. Some authors have attempted to discriminate the target from background regions [294–296]. Furthermore, motion can be leveraged using image registration [42, 297]. Another strategy is based on similarity comparisons between the target object and proposals from the search image [289–292]. In this latter case, Siamese networks are the most popular choice to extract and compare deep features from the images [128]. Recently, attention-based models have been adapted for object detection and tracking. For instance, Detection Transformer (DETR) [127] relaxed the dependence on region proposal by inferring an arbitrary number of object queries using a Transformer backbone. This work was extended for object tracking [293]. Also, the Transformer have been inserted into a Siamese-like model for video object tracking [298]. Generally, for medical datasets, tracking models receive a template image, with the target landmark, and a search image in which to estimate the landmark position [128]. However, this task becomes more challenging when it involves different image dimensionalities.

## 9.3 Preliminaries

### 9.3.1 Problem formulation

Given the input images $\langle I_t, I_{t-1}, \dots, I_{t-m} \rangle$ at $m$ observed time steps, the first goal is to produce visual representations corresponding to $n$ future times $\langle z_{t+1}, z_{t+2}, \dots, z_{t+n} \rangle$ which can be used to generate either volumetric deformations, if used with a motion model, or in-plane deformations corresponding to future times. By warping a reference image with the predicted deformations we can yield future images $\langle I_{t+1}, I_{t+2}, \dots, I_{t+n} \rangle$. The second goal is to learn a non-linear parametric mapping from the outcomes of a deep motion model to refined deformations over a region of interest $\left\langle \phi_{t+1}^{ROI}, \phi_{t+2}^{ROI}, \dots, \phi_{t+n}^{ROI} \right\rangle$.

### 9.3.2 Attention mechanism

The Transformer was originally proposed as a new paradigm for machine translation [7] since, in contrast to previous models, no convolution operation nor recurrence was involved in the computations. Since its introduction, it has shown to be a powerful mechanism to process sequences and has gained increasing popularity in many tasks. Transformer can attend to complete sequences thereby aggregating information amongst the inputs and learning long-range relationships. The key concept behind Transformers is the scaled dot-product attention mechanism, where the input is linearly projected to a set of queries $Q \in \mathcal{R}^{n_q \times d_q}$, keys $K \in \mathcal{R}^{n_k \times d_k}$, and values $V \in \mathcal{R}^{n_v \times d_v}$. The vector dimensionality $d_q$ equals to $d_k$, the number of keys $n_k$ equals to the number of values $n_v$. The output of the attention layer is given by computing the weighted sum of the values, where attention scores $S \in \mathcal{R}^{n_q \times n_k}$ are calculated from the queries and key as follows:

$$A(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9.1}$$

where the $Softmax$ function is used to normalize the scaled dot-product attention scores.

### 9.3.3 4D deep motion modelling framework

The conditional motion model proposed in [191] receives as input a reference volume $V_{ref}$, acquired at a fixed respiratory phase, an image sequence $I_{seq} = \langle I_t, I_{t-1}, \ldots, I_{t-m}\rangle$, and a set of temporal volumes $\langle V_{t+1}, V_{t+2}, \ldots, V_{t+n}\rangle$ for their creation (i.e. training stage). The goal of the model is to relate the 2D slices at $m$ previous times, with high-dimensional deformations within a latent space. Furthermore, the model contains a temporal predictor which yields extrapolated-in-time visual representations from the input slices. Nonetheless, the motion model is agnostic to the approach used for visual representation forecasting. During testing, only $V_{ref}$ and the cine-acquisitions are available. In consequence, the model relies mostly on the visual representations corresponding to future times. Thus, the temporal prediction of accurate and meaningful representations is crucial for the model performance.

## 9.4 Proposed method

In this section, we describe in details the proposed prior-based Transformer architecture for temporal prediction as well as the phase-conditioned tracker module. Figure 9.1 depicts an overview of the entire framework. The blue boxes represent the two blocks containing the contributions presented in this work. Our framework assumes a motion model, as described in Section 9.3.3, which receives forecasted visual representations from the temporal predictor. Furthermore, these ahead-of-time representations are fed, together with the displacement vector fields (DVF) generated by the motion model, to the tracker module to yield refined

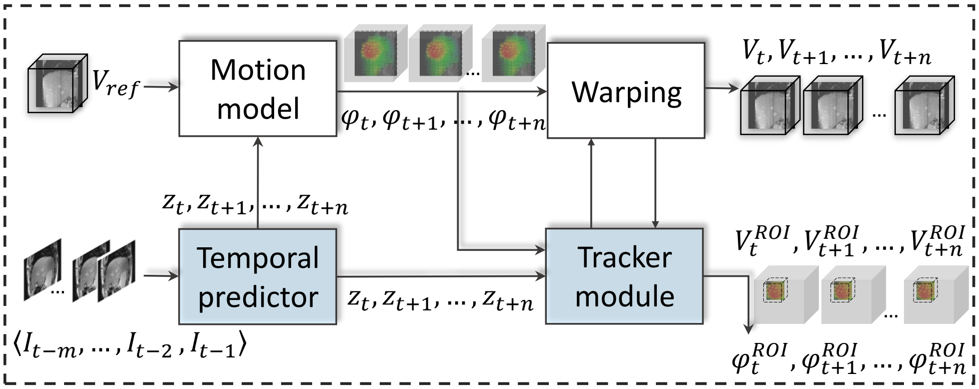

$V_{ref}$
Motion model
Warping
$V_t, V_{t+1}, \dots, V_{t+n}$
$\varphi_t, \varphi_{t+1}, \dots, \varphi_{t+n}$
$z_t, z_{t+1}, \dots, z_{t+n}$
Temporal predictor
$z_t, z_{t+1}, \dots, z_{t+n}$
Tracker module
$V_t^{ROI}, V_{t+1}^{ROI}, \dots, V_{t+n}^{ROI}$
$\langle I_{t-m}, \dots, I_{t-2}, I_{t-1} \rangle$
$\varphi_t^{ROI}, \varphi_{t+1}^{ROI}, \dots, \varphi_{t+n}^{ROI}$

Figure 9.1 Framework overview. The temporal predictor block provides ahead-of-time visual representations, which act as predictive variables for the motion model. The outcome of the temporal predictor and the motion model is used to refine the deformation over a pre-selected region of interest.

motion fields over a region of interest (ROI).

### 9.4.1 Temporal predictor

Figure 9.2 shows a schematic representation of the proposed temporal predictor, which is composed by: (1) a CNN-based backbone for feature extraction, (2) an encoder-decoder Transformer architecture with learnable queries, (3) attention-based encoders to learn a prior from ground-truth images, and (4) a CNN-based feature projector. The latter component is optional since it depends on the specific application. For instance, for the framework presented in Figure 9.1 this block is not required. On the other hand, it could be added to enable the prediction of 2D+$t$ deformations.

#### Frame-wise feature extraction

The frame feature encoder receives an input sequence containing the channel-wise concatenation (denoted as $|$) of temporal images with a fixed reference image. It is shared by all frames and can be implemented by any CNN or self-attention-based models. During training, it generates a feature map sequence $Z_p \in \mathcal{R}^{m \times H \times W \times C}$ for $m$ past frames $\{I_t|I_{ref}, I_{t-1}|I_{ref}, \dots, I_{t-m}|I_{ref}\}$ where $I_i|I_{ref} \in \mathcal{R}^{H_0 \times W_0 \times C_0}$, $H_0$, $W_0$ and $C_0$ and $H$, $W$ and $C$ denote the initial and final image height, width, and channels, respectively. Likewise, this block receives the ground-truth image sequence $\{I_{t+1}|I_{ref}, I_{t+2}|I_{ref}, \dots, I_{t+n}|I_{ref}\}$ ($n$ is the number of future frames) from which computes a feature map sequence $Z_f$ to be used as prior knowledge (see Section 9.4.1). During inference, only the past frames are fed into this

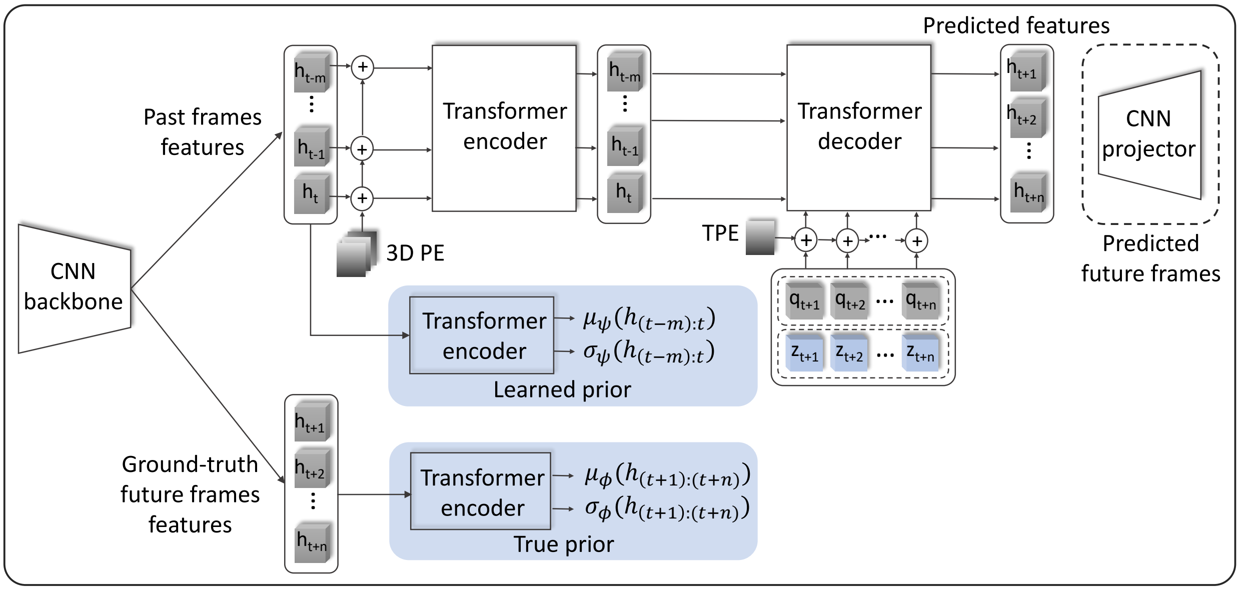


Figure 9.2 Schematic representation of the proposed temporal predictor. During the training stage, it receives an input sequence containing both past and ground-truth future frames. The features extracted from this input sequence are used to learn prior knowledge, which is combined with the queries of the Transformer decoder to forecast visual representations associated to future times. The projector block recovers dense deformations from the forecasted features.

block.

### Transformer with learnable queries

We employ the powerful Transformers as backbones for our temporal predictive model. Since the encoder transformer expects a sequence as input, the previous frame features need to be flattened over both temporal and spatial dimensions to get a sequence of uni-dimensional vectors, i.e. $Z_p \rightarrow Z_p \in \mathcal{R}^{(m \times H \times W) \times d_{model}}$. Furthermore, since the transformer architecture is permutation invariant, it is required to incorporate the position information for each temporal feature maps using fixed 3D positional encodings. The positional encoding for a pixel position $(t, h, w)$, where $t \in [0, m-1]$, $h \in [0, H-1]$, $w \in [0, W-1]$, is a 1D vector with the same dimensionality as the Transformer ($d_{model}$). It is formed by concatenating three 1D positional encodings along each dimension. Thus, $d_{model}$ is constrained to be divisible by 3. Positional encodings ($PE$) of each dimension are encoded independently by sine and cosine functions of different frequencies:

$$PE(pos, i) = \begin{cases} sin(pos/\omega_k), \text{for } i = 2k \\ cos(pos/\omega_k), \text{for } i = 2k+1 \end{cases} \tag{9.2}$$

where $pos$ denotes position, $i$ is encoding dimension, and $i \in [0, d_{model}/3 - 1]$. $\omega_k = 10000^{2k/d_{model}/3}$. Subsequently, they are added to the visual feature sequence. We choose the 3D positional encodings to be fixed rather than learnable as previous studies [7,127] have shown that both approaches yield similar results. In addition, the learnable variant would increase the number of model parameters.

Given $Z_p \in \mathcal{R}^{(m \times H \times W) \times d_{model}}$, the Transformer encoder makes every element of $Z_p$ attend to each other. It follows the original architecture [7] meaning that queries, values and keys are all derived from the same inputs aggregating the contextual information from the sequence itself. The Transformer decoder aims to decode features of each future frame based on the encoder outputs and future frame queries. Sequence of frame queries $\langle q_{t+1}, q_{t+2}, \ldots, q_{t+n} \rangle$ are initialized and fed into the frame-level decoder, where $q_i \in \mathcal{R}^{d_{model}}$ denotes the query corresponding to the features of the $i^{th}$ predicted frame. In contrast to previous works [127,128], we integrate an additional temporal positional encoding for the deformation queries in order to maintain the order of predicted future frames. The temporal positional encoding (TPE) is implemented as Equation 9.2, except that $i \in [0, d_{model} - 1]$ and $\omega_k = 10000^{2k/d_{model}}$. Future deformation queries are learned automatically during training. Output features are used for the following future deformation generation.

### Prior-based latent modeling

Considering the future frame prediction to be a generative task, inspired by [144], we propose to improve the regression of future visual representations by providing an additional source of information using an attention-based prior. The aims of the learnable prior are twofold. Firstly, it leverages additional information available during training. Secondly, it acts as an stochastic regularizer during the learning of the queries. In contrast to [144], $n$ prior distributions $p_\theta(h_{t+i}|z_{t-m:t+i}) \quad i = 0, 1, \ldots, n$ are estimated simultaneously from the sequence of feature vectors extracted from the ground-truth images up to a future time $t + i$. Moreover they are parameterized by $\mu$ and $\sigma$, i.e. $\mathcal{N}\left(\mu(z_{t-m:t+i}), \, \sigma(z_{t-m:t+i})\right)$. In parallel, $n$ identical networks try to learn an approximation of the prior distributions but using limited observations, i.e. $r_\psi(h_{t+i}|z_{t-m:t-1}) \quad i = 0, 1, \ldots, n$. In other words, these approximations networks aim at learning a mapping function between the spatiotemporal information contained on the observed frames and the latent variable describing a future time. This is performed by enforcing both distributions $p_\theta$ and $r_\psi$ to be close each other by minimizing a KL-divergence term:

$$\mathcal{L}_{KL} = \sum_{i=0}^{n} KL \left[ p_\theta(h_{t+i}|z_{t-m:t+i}) | r_\psi(h_{t+i}|z_{t-m:t-1}) \right] \tag{9.3}$$

During training, latent variables $h_{t+i}$ are sampled from $p_\theta$ and concatenated with the queries $q_{t+i}$ $(i = 0, 1, \ldots, n)$. At test time, the sampling is performed over $r_\psi$.

The distance between distributions (Eq. 9.3) is inserted within the total loss function, which also aims at minimizing the reconstruction loss:

$$\arg\min_\theta \left[ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{KL} + \mathcal{L}_{rec} \right] \tag{9.4}$$

The specific formulation of the reconstruction term depends on whether the temporal predictor is used within a 4D motion model or for dynamic slice forecasting.

### Feature projector

Given the predicted future frame features from the Transformer decoder, the feature projector aims at reconstructing future dense deformations. This block can be implemented by a standard deconvolution neural network and shared by each future frame features. The last layer should output a 2D motion field. Since we avoid direct pixel synthesis, the motion fields generated by the feature projector are used to resample (operation denoted with $\circ$) the reference image using a spatial transformation layer, thereby yielding $n$ forecasted images. This enables to calculate a reconstruction term ($\mathcal{L}_{rec}$), comprised by the similarity between predicted and ground-truth images and a diffusion regularizer on the spatial gradients to ensure smooth motion fields:

$$\mathcal{L}_{rec} = \frac{1}{n} \sum_{i=1}^{n} \left[ \mathcal{L}_{sim}\left(I_i, \hat{I}_i\right) + \alpha \mathcal{L}_{smooth}\left(\hat{\phi}_i\right) \right] \tag{9.5}$$

where $\hat{I}_i$ results from warping $I_{ref}$ with the estimated motion $\hat{\phi}_i$, $\alpha$ is a regularization parameter, and $\mathcal{L}_{smooth}(\hat{\phi}_i) = \sum_{p \in \mathbb{R}^2} \|\nabla \phi\,(p)\|^2$ computes the differences between neighboring 2D positions $p$.

## 9.4.2 Model-based tracker module

In current frameworks designed for target tracking, such as [287], it is assumed that both source and target volumes are available. However, this assumption is not met during real-time image-guided interventions. We tackle this shortcoming by leveraging the global estimation provided by a motion model to better localize the target. In [190, 191], it was shown that the latent space of autoencoding-based motion models carries important information related to the respiratory phase. Ultimately, free-breathing is the underlying factor of variation. Such low-dimensional representations are uncovered thanks to the capacity of autoencoders to learn a non-linear parametric mapping from volume deformations to their latent vectors. On the other hand, it is important to mention that generating 3D deformations from 2D images is an ill-posed problem.

Considering these elements, we propose to learn a mapping function, conditioned on latent vectors, to refine the target trajectory from the coarse approximation of the entire field-of-view provided by the model. The core idea is to use the latent codes of the motion model to compute an attention map over the coarse deformation fields. We assume that the target position is defined a priori. In the context of image-guided radiotherapy this is feasible since a pre-operative (fixed) volume is routinely acquired before the procedure.

Given the selected target position $(x_{ref}, y_{ref}, z_{ref})$, a three-dimensional bounding box can be defined with origin at $(x_{ref} - \frac{\Delta_x}{2}, y_{ref} - \frac{\Delta_y}{2}, z_{ref} + -\frac{\Delta_z}{2})$ and height, width and depth denoted by $\Delta_x$, $\Delta_y$, $\Delta_z$, respectively. We use the bounding box to mask the motion model's outcome. Then we divide the motion components ($\phi_x^{ROI}$, $\phi_y^{ROI}$, $\phi_z^{ROI}$) and compute weighted maps ($S_x$, $S_y$, $S_z$) for each one. The refined 3D displacement at an arbitrary time ($\hat{\phi}_i^{ROI}$), corresponding to the motion plane $i$, is given by the concatenation of the element-wise multiplications of the motion model prediction ($\phi_i^{ROI}$) and the weighting coefficients ($S$), i.e., $\hat{\phi}^{ROI} = Concat(S_x \times \phi_x^{ROI}, S_y \times \phi_y^{ROI}, S_z \times \phi_z^{ROI})$. The weighted map at a given plane $i$ is computed following [299]:

$$S_i = \sigma_2(\sigma_1(W_c c + W_\phi \phi_i^{ROI}) W_s) \tag{9.6}$$

where $c$ is the latent vector from the motion model, $\sigma_1$ and $\sigma_2$ are ReLU and sigmoid activations, and $W_c$, $W_\phi$ and $W_s$ are linear transformations implemented with $(1 \times 1 \times 1)$ convolutions.

During the training of the tracker module, the optimization problem aimed at minimizing the dissimilarity between the ground-truth and predicted patches as well as the difference between ground-truth and predicted motion fields.

## 9.5 Results

In this section, we demonstrate the effectiveness of our model both for temporal forecasting and tracking on an MRI dataset containing 4D volumes acquired on 25 subjects. Details on the acquisition parameters, reconstruction and pre-processing of this dataset can be found in [191]. First, we provide the implementation details of the models as well as the validation methodology. Next, we show results on future slice generation and model-based volume generation. In addition, we compare our approach with related state-of-the-art predictive models. Finally, we present the results of the tracker module.

### 9.5.1 Implementation details

**Temporal predictor** The feature encoder of the bi-channel image sequence was implemented as described in [191]. All the involved multi-head attention blocks were composed by one layer, 16 heads, 2048 channels in the internal feed-forward layer and Dropout of 0.1. Layer normalization was placed before computing attention, as explained in [300]. The sub-networks involved in the prior computation end in two 256−sized fully connected layers, which determine the parameters of the latent distribution. The projector network was composed of successive pairs of 2D transposed convolution and convolution layers with kernel size $3 \times 3$ to reach a final size of $2 \times 64 \times 64$. Each layer was followed by LeakyReLU activation set to 0.2.

**Tracker** The right portal vein was annotated in the reference volumes by a radiologist. Although the landmark represents the same anatomical structure, it has a variable appearance across the subjects, as can be observed in Figure 9.3. A bounding box with dimension $4 \times 8 \times 8$, and centered at the annotated position, was selected as region of interest. During the optimization of the tracker, the weights of the motion models remained unchanged. The hidden dimension of the tracker module was set to 32.

**Training details** The network's parameters were optimized using the Adam optimizer with an initial learning rate ($lr$) set at $10^{-4}$. Training was performed in PyTorch with a batch size of 10. We used a leave-one-out validation scheme on a subject level. In equation 9.5, the negative local cross correlation (NCC) was used as similarity measure and $\alpha$ was set to 0.0001.

### 9.5.2 Validation methodology

We assess the model performance using different measures based on motion and appearance. Specifically, we report the geometrical error for all the respiratory states, which is defined

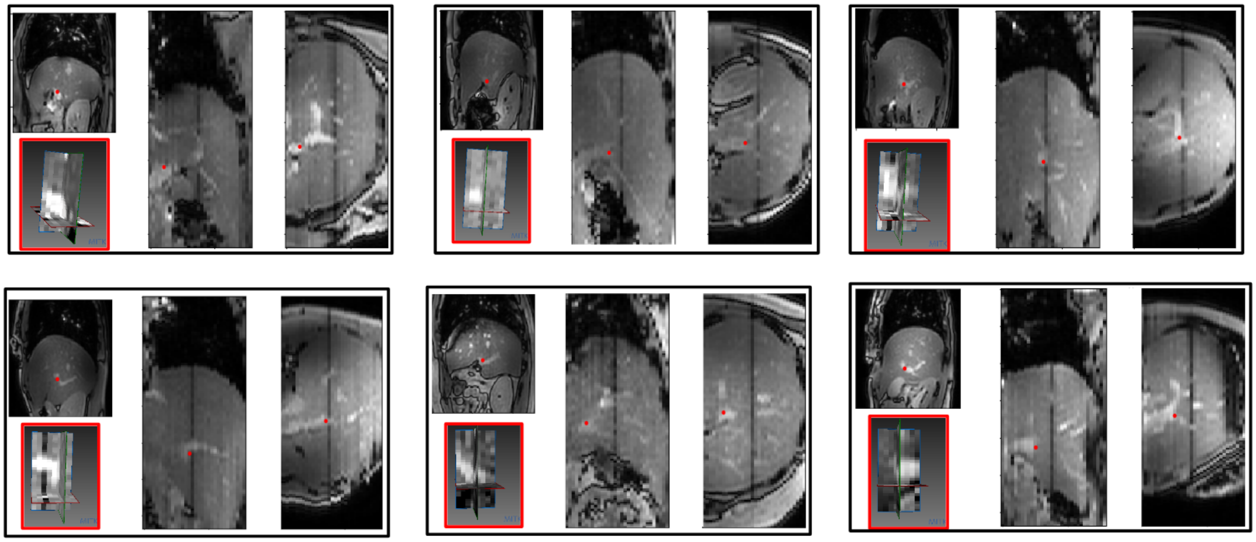


Figure 9.3 Examples of landmarks for tracking. The annotated position of the right portal vein is represented by the red point at different anatomical orientations.

as the Euclidean norm of the voxel-wise vector difference between the ground truth and predicted motion fields. Notwithstanding the ground-truth motion does not necessarily represent the real motion due to errors introduced by the registration process, it still represents a valid reference [191]. Additionally, we used similarity metrics such as MSE, NCC, Structural Similarity (SSIM) and Peak Signal-to-Noise Ratio (PSNR), which are generally used in related works to measure the spatial consistency of the predicted volumes. Statistical significance was computed using a Wilcoxon signed-rank test, considering $p < 0.01$ to indicate a significant difference. Effect size was measured using Pearson correlation coefficient ($\rho$). Finally, we followed a leave-one-out scheme, which means results are reported on unseen test cases.

Table 9.1 Geometrical errors (in mm) and image similarity measures obtained with the motion model when using different temporal predictors. Values are mean ± std ($95^{th}$percentile).

| Predictive module | TRE (mm) | NCC | MSE |
|---|---|---|---|
| ConvGRU [190] | $1.60 \pm 1.09$ (3.17) | $0.71 \pm 0.14$ (0.89) | $0.16 \pm 0.09$ (0.32) |
| ConvLSTM [191] | $1.37 \pm 0.92$ (2.60) | $0.76 \pm 0.13$ (0.91) | $0.13 \pm 0.09$ (0.22) |
| Transformer | $1.34 \pm 0.87$ (2.51) | $0.76 \pm 0.13$ (0.91) | $0.14 \pm 0.09$ (0.23) |
| **Transformer+prior** | $\mathbf{1.25 \pm 0.74(2.13)}$ | $\mathbf{0.81 \pm 0.11(0.95)}$ | $\mathbf{0.10 \pm 0.07(0.18)}$ |

### 9.5.3 Model-based future volume generation

Table 9.1 presents the geometrical error (in mm) measured between ground-truth motion fields and those predicted by the motion model. It also contains similarity metrics between ground-truth and predicted volumes for a predictive horizon of 450 ms. Moreover, it shows a comparison between different predictive mechanisms, namely, ConvGRU and ConvLSTM, which were used in the motion models described in [190] and [191]. In addition, a Transformer model with learnable queries, and the proposed model which also includes the prior. The models were trained under the same conditions using coronal orientation for the surrogate images. Also, the same network architecture was used for feature extraction.

Overall, the proposed model yields the lowest geometrical error compared to the other variants, a result that was found to be statistically significant, $p \ll 0.01$, $\rho = 0.48$ (proposed/ConvGRU), $\rho = 0.73$ (proposed/ConvLSTM), $\rho = 0.72$ (proposed/Transformer). Likewise, the MSE and NCC values show a similar behavior.

### 9.5.4 Temporal forecasting of in-plane deformations

The performance of the proposed model was confronted to several state-of-the-art approaches for future image generation. We used the publicly available implementations released by the authors. The first approach, introduced for stochastic video generation (SVG-LP), computes a prior from the next image recursively using an LSTM [144]. For the frame encoder and decoder, we used the VGG16 [301] architecture as described in [144]. The second model [119] consists of an U-net-like shape with ConvLSTM. Furthermore, it was designed specifically for respiratory-induced deformation forecasting and integrates spatial transformations. The third model leverages ground-truth images during the training stage to learn a long-term motion context memory (LMC-Memory) with memory alignment. Except [119], these models were validated originally with natural images. Furthermore, they were designed to directly synthesize the pixels.

Table 9.2 reports similarity metrics (for the next time step) between ground-truth and predicted images for the proposed model as well as for comparative approaches. It also present the computational time required for inference when deployed on a NVIDIA Titan RTX GPU with 64 Gb RAM. The presented values were obtained by averaging 50 measurements. Overall, the proposed model obtained the best performance in all the presented metrics. In terms of computational time, all the methods are quite similar except LMC-Memory, which got a slightly higher value.

Figure 9.4 illustrates visual results of the generated images by each implemented method. The

Table 9.2 Comparison with the state-of-the-art methods. Values are mean ± std.

| Method | PSNR | SSIM | MSE | Time (ms) |
|---|---|---|---|---|
| SVG-LP [144] | $17.35 \pm 1.14$ | $0.59 \pm 0.08$ | $0.12 \pm 0.03$ | $\mathbf{7.46 \pm 0.36}$ |
| Recurrent U-net [119] | $25.48 \pm 5.52$ | $0.77 \pm 0.13$ | $0.08 \pm 0.04$ | $8.12 \pm 0.15$ |
| LMC-Memory [302] | $23.55 \pm 2.37$ | $0.71 \pm 0.13$ | $0.10 \pm 0.04$ | $11.20 \pm 1.05$ |
| **Proposed** | $\mathbf{26.30 \pm 4.55}$ | $\mathbf{0.78 \pm 0.11}$ | $\mathbf{0.07 \pm 0.05}$ | $7.78 \pm 0.21$ |

images cover a respiratory cycle. It can be observed that deformations near the inhale phase are the most challenging regardless the method. Furthermore, considering the difference maps, the proposed method achieves the best spatiotemporal consistency.

### 9.5.5 Model-based tracker results

Table 9.3 presents the target registration errors on the selected region of interest for different model variants, which were trained under the same conditions to predict the next three temporal volumes. Since the volumes has a temporal resolution of 450 ms, the predictive horizon is 1350 ms. The first row contains the errors when there is no motion compensation, i.e. the initial motion. To quantitatively analyze the effectiveness of the tracker module, we compare the results before and after its integration to the motion model. Moreover, several variants of motion models were considered, i.e. using different temporal predictors. Experimental results show that, in average, the tracker module can decrease the alignment errors by 63% when compared to the motion model outcome. Also, it is important to mention that the tracking areas are out-of-plane with respect to the position of the input images. Figure 9.5 depicts exemplary sagittal and coronal slices, taken at the center of the tracked region, in the predicted volumes. It can be observed the error reduction in the results obtained with the tracker module.

## 9.6 Discussion and conclusion

In this work, we presented a novel self-supervised model able to predict future representations from an image sequence by learning queries within a Transformer architecture. Alternatively to previous works, based on autoregressive predictions with recurrent networks [119, 144], our approach enables parallel decoding. Furthermore, the number of learnable queries is defined by the number of future time steps. For certain applications, parallel decoding is

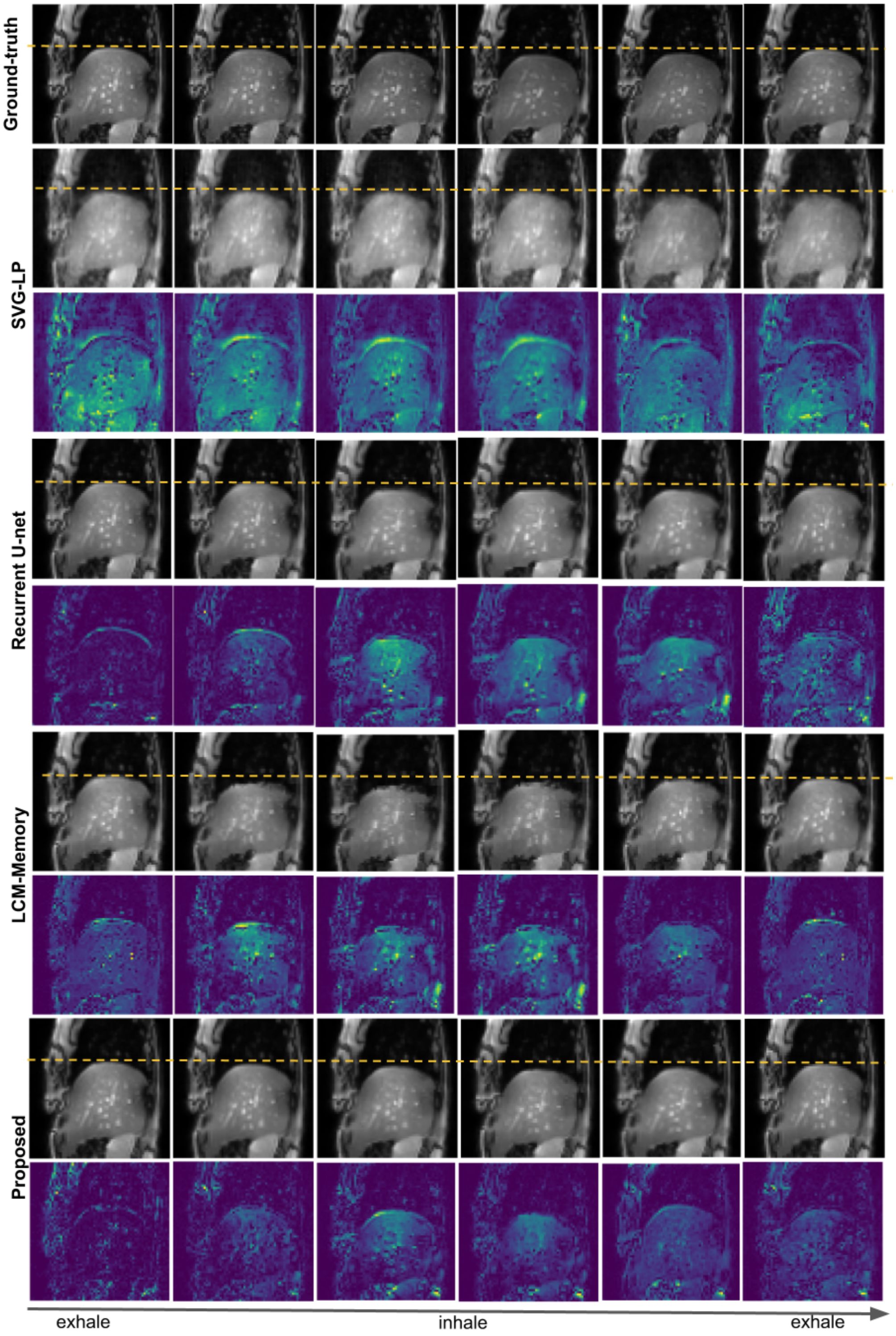


Figure 9.4 Visual results. The top row contains the ground-truth images spanning a respiratory cycle. The second, fourth, sixth and eight rows show the predictions performed by the implemented methods and at the bottom the difference maps.

Table 9.3 Target tracking errors (in mm) measured at selected respiratory phases for the V-MRI dataset. These values were measured for the next time step, i.e. a horizon of 450 ms. Overall values consider all the phases. Values are mean ± std [$P_{90}$].

| Method | TRE (450 ms) | TRE (900 ms) | TRE (1350 ms) |
|---|---|---|---|
| Initial motion | 6.52 ± 3.41 [8.19] | 6.35 ± 3.11 [8.0] | 6.42 ± 3.40 [8.23] |
| MM + GRU | 2.65 ± 1.93 [5.47] | 2.72 ± 1.89 [5.45] | 2.66 ± 1.86 [5.38] |
| MM + LSTM | 2.68 ± 1.73 [4.90] | 2.66 ± 1.70 [4.81] | 2.59 ± 1.66 [4.68] |
| MM + Transf. | 2.61 ± 1.58 [4.80] | 2.56 ± 1.57 [4.73] | 2.54 ± 1.55 [4.67] |
| MM + Transf. + prior | 2.55 ± 2.11 [6.22] | 2.56 ± 1.45 [4.70] | 2.60 ± 2.08 [6.15] |
| MM + GRU + tracker | 1.75 ± 1.19 [3.17] | 1.78 ± 1.19 [3.19] | 1.77 ± 1.17 [3.13] |
| MM + LSTM + tracker | 1.66 ± 1.21 [3.25] | 1.61 ± 1.16 [3.13] | 1.57 ± 1.13 [3.03] |
| MM + Transf. + tracker | 1.65 ± 1.17 [3.21] | 1.63 ± 1.16 [3.16] | 1.61 ± 1.15 [3.11] |
| **MM + Transf. + prior + tracker** | **1.56 ± 1.13 [3.09]** | **1.53 ± 1.11 [3.04]** | **1.52 ± 1.10 [2.98]** |

more convenient than the autoregressive manner. For instance, it could be useful to track trajectories or to assess the spatiotemporal consistency across consecutive samples.

Our model combines the powerful visual feature extraction capability of convolutional neural networks with the strong representation capacity of Transformers. The visual representations obtained from the images are employed as tokens for the Transformer. Due to the sequential nature of cine acquisitions, this attention-based model is inherently well suited for this predictive task.

Inspired by [144], we learn a prior from ground-truth images, which are available during the training stage. Nevertheless, in contrast to that work, we followed a multi-time approach thereby avoiding the auto-regression, which is prone to error propagation. Additionally, unlike [144], this supplementary knowledge was not combined with the image features but with the Transformer queries, which contain the actual predictions. Results showed that, conditioning the queries to the prior information yields improved results, which can be attributed to narrowing the prediction space.

On the other hand, many state-of-the-art techniques employ skip connections [119, 144]. For instance, in object-centric datasets it helps to the generation of static background features [144]. While skip connections have been shown to boost the generation quality, they impose a dependency between encoding and decoding parts. Since we designed our temporal predictor to be integrated within a motion model, we do not rely on skip connections. In the case of such integration, the data dimensionality managed by the temporal predictor will differ from the one in the motion model. Thus, we rather let the deformation regression to be fully generative. Nonetheless, alternative ways to tackle the dimension disparity and to leverage skip connections should be explored in future works.

Although the proposed temporal predictor can be used primarily to regress future images, we also demonstrated how the forecasted representations can be integrated within a motion model to allow future volume generation. This integration led to improved results with respect to previous outcomes. This is very likely due to direct access that each input sample has to all the other inputs, which prevents information loss. Likewise, for future frame generation it yielded improved results compared to state-of-the-art techniques. Experiments with comparative approaches evidenced that, given the complexity of medical datasets, using spatial transformations is more feasible than regressing pixel intensities and led to sharper results.

The main limitation of image generation based on spatial transformations is the incapacity to maintain structures that are not contained in the source image. However, for the analysis of motion in medical datasets, e.g. deformable registration, it is common to measure and apply the deformation using a source image. Our experiments also confirmed that there is a huge difference between object-centric representations and medical datasets, where typically is harder to separate the background and the foreground. Hence, methods with excellent performance in the former case may face obstacles to predict the whole scene. Given the high dimensionality and complexity of these challenging datasets, models are required to build a deep understanding of the underlying process.

We also introduced a tracker module, which leverages the outcome of a 4D motion model to refine the deformation fields within a pre-selected region containing the target. We addressed such a task by using the temporal latent representations as gating signal to refine the displacement fields. Experimental results showed that the error is consistently reduced regardless of the nature of the temporal predictor. Therefore, this approach enables 3D target localization from 2D slices. Furthermore, it represent a valid alternative for real-time image-guided interventions, where pairs of up-to-date volumes are not available.

Since our wok is targeted at deformation forecasting, we consider inputs to be a concatenation of both source and template images. In our experiment, the source image (also known as reference image in this manuscript) was extracted from the same dataset. Thus, it has the same acquisition parameters as the rest of the images. However, in practice, this assumption may not be fulfilled. Therefore, we identify it as a potential limitation that deserves further validation.

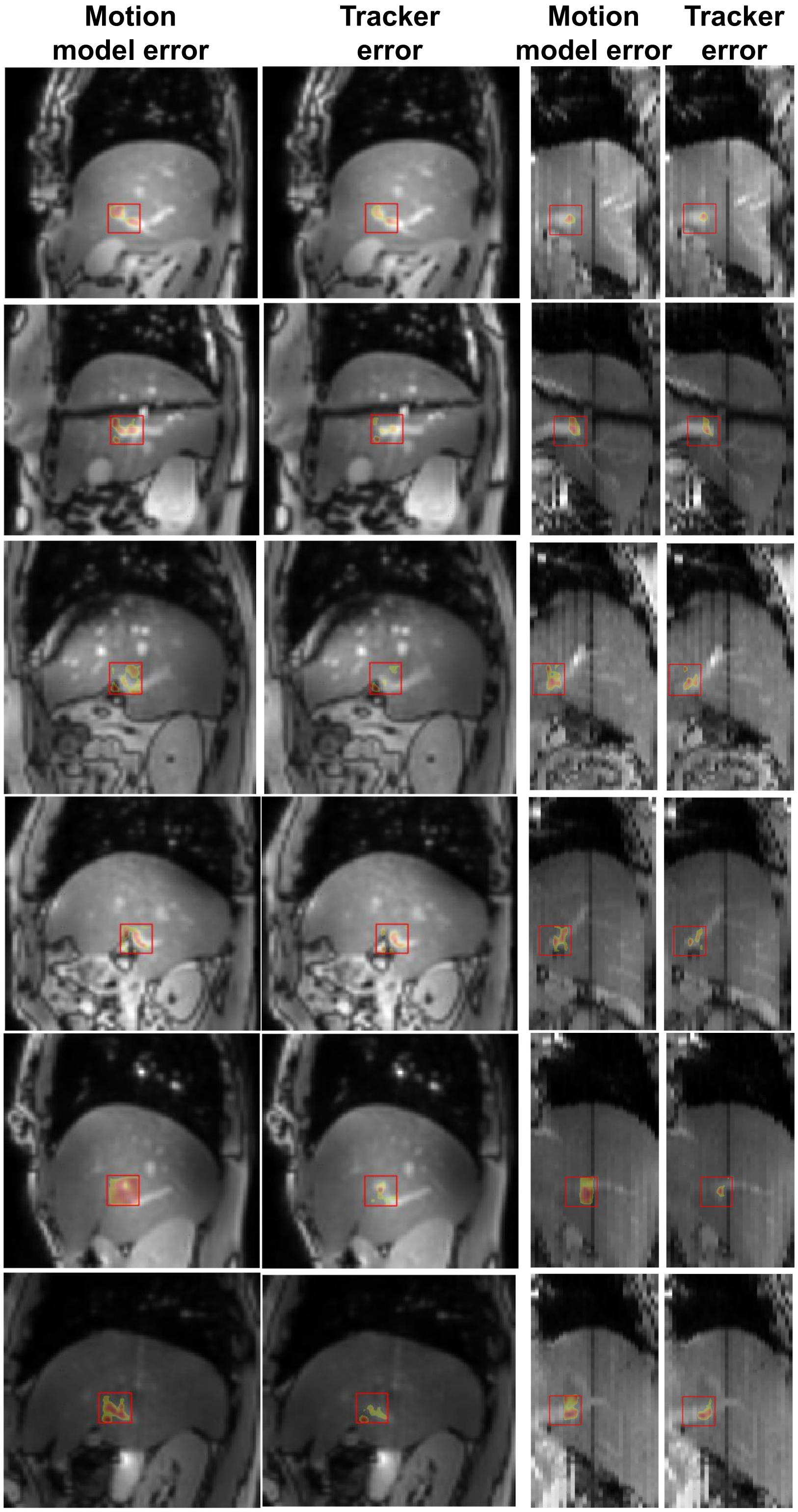


Figure 9.5 Visual examples of the error reduction when using the tracker module. The red box shows the tracker region.

# CHAPTER 10 GENERAL DISCUSSION

The general methodology in this thesis was guided by three research objectives, which led to various novel solutions for imaging, modelling, and analysis of respiratory motion. Firstly, an automated method was developed in order to construct temporal volumes from navigator-less cine acquisitions. Secondly, two frameworks were proposed to relate partial observations with high-dimensional deformations. Moreover, both deterministic and probabilistic approaches were explored. Finally, methods for future image forecasting and target tracking were designed by leveraging attention structures. The development and advantages brought by these new methodologies will be discussed in this chapter.

## 10.1 4D image formation

The construction of 4D datasets is an important step to observe and quantify the dynamic behavior of moving organs. Our proposed method for 4D image formation considers the use of MRI-based imaging technology given their well-known advantages. Specifically, it provides excellent soft tissue contrast, emits no ionizing radiation exposure, and is flexible in selecting image plane position and orientation. Indeed, this modality is highly desirable in several clinical applications. The development of this method was motivated by several current shortcomings. For instance, respiratory surrogates are crucial during the sorting process from multi-slice cine acquisitions. However, it is not uncommon clinical scanners do not offer the possibility to acquire these complex sequences. On the other hand, many self-gating strategies are not fully automatic or cannot deal with subjects with complex respiration cycles. In fact, for some patients it becomes a major hurdle.

Obtaining a navigator signal from dynamic slices may be a challenging process. One major limitation is the detection of reference phases, which at the time of publication was performed manually. Thus, perhaps the biggest contributions in our methodology are the automatic extraction of the breathing signal and the end-exhale phase detection. We have shown that, compared to other methods assuming a regular cycle, our method is able to cope with irregular breathing and with small apneas. This is, in part, due to the robustness of these two steps.

On the other hand, the graph-based approach enables an arbitrary slice selection to start the shortest-path computation. In theory, such flexibility allows the reconstruction of different 4D image sequences. Hence, this increases the variability of respiratory motion patterns that an operator can choose to capture, which is especially interesting for motion modelling.

Moreover, it could be used as an alternative strategy for data augmentation.

Experimental analysis revealed that this method is suitable to work with high spatial and temporal resolution data. During our experiments, we noticed that manifold learning based methods are susceptible to outliers points. One solution for this could be filtering the abnormal cycles before alignment. The volunteers that participated in the acquisition protocol were instructed to breath normally. However, when processing the acquired images we observed that, during the same acquisition, subjects can present very variable breathing patterns. For instance, we observed cycles containing deep breathing, prolonged periods of holding breath, and very shallow breathing. These cases make the slice reordering even more challenging. For instance, when slices are acquired under these conditions, they often appear as outliers in the low-dimensional space used by manifold theory. In summary, any robust 4D construction method should be able to cope with these outlier cycles automatically.

The 4D image formation method presented in Chapter 5 exhibit some desirable properties. First, it is fully automatic while not sacrificing robustness to other impediments such as irregular breathing patterns. Second, it can be generalized. This means that, with minor changes mainly relating to the spatial distribution of the organ within the image, it can be applied to other moving organs such as upper airways, heart and other abdominopelvic structures. Third, the 4D construction principle is independent of the number of acquired slices across the organ. This implies that it can be employed even when imaging is done to gather only partial data instead of the full set across the organ of interest.

The main limitation of the proposed method, which is common for all the related approaches relying on multi-slice, is that physiological correctness is difficult to ensure, even if temporal coherence is achieved. Furthermore, we observed that slice stacking at the inhalation state is more difficult because the liver does not always descend to the same position. To accurately reconstruct the area of interest, it is important to ensure that most of the possible combinations of respiratory motions are acquired for all the slices.

## 10.2 4D motion modelling from image surrogates

Certainly, obtaining quality motion data from the 4D observations is a key step before building a motion model. Generally, one is interested in modeling the deformations undergone by the organ between a reference respiratory phase and other phases. The deformation between a pair of volumes is generally computed using deformable image registration. Over the years, an enormous amount of research has been dedicated on this particular topic. In this project, proposing new techniques to perform this task was not explored within the scope of the research. In fact, the developed solutions are agnostic to the method used for motion estimation.

In Chapters 6 and 7, we developed motion models used deterministic and probabilistic approaches, respectively. In contrast with classical approaches, these models rely exclusively on deep neural networks. One common aspect amongst both solutions is that dimensionality reduction, via autoencoding, is the basic principle for motion modelling. The autoencoding process aims at producing an efficient compressed representation that enables input reconstruction. Such compressed representation is a key attribute since it avoids learning an identity mapping between the input and output. Therefore, it must contain relevant information to allow the decoder maps from the low-dimensional space back into the original space.

The model described in Chapter 6 is designed to create low-dimensional manifold representations of 3D non-rigid deformations which are associated with surrogates images. It benefits from the spatiotemporal information contained in the dynamic 2D slices, which are also mapped to a low-dimensional space. The association between the 3D deformations and the surrogates is done by minimizing the L2 distance between representations. This model is based on convolutional autoencoders and does not involve any stochastic component. Alternatively, the model of Chapter 7 presents a probabilistic formulation for the motion modelling task. It is based on conditional variational autoencoders, which are generative models able to learn a probability distribution conditioned on certain variables. We found this conditioning as a plausible solution for the task in question, i.e., relating deformations with partial observations. Hence, we formulate a conditional manifold learning task to relate the predictive variables with their corresponding deformation encoding.

Similarly to the previous model, the backbone consists of a probabilistic autoencoding process that learns how to compress and recover the input 3D deformations while conditioning the generation on respiratory phases. We have shown that, with the autoencoding approach, similar data points are mapped close to each other in the latent space according to their respiratory phase. This feature enforces the model's interpretability. Furthermore, it could be potentially used for further classification or other downstream tasks.

Another interesting characteristic is the use of a static reference volume taken at end-exhale as additional source of information and conditioning variable. Since the 3D generation from 2D slices is an ill-posed problem, the developed models leverage the volumetric information provided by the reference volume. Generally, a volume gated at a fixed respiratory phase is routinely acquired before treatment. This volume can be used as reference to obtain the deformations. During our experiments, it was extracted from the 4D dataset. However, in a real-life situation the contrast may differ from the one used to train the model. We hypothesise that this difference should not represent an obstacle as long as the following two premises are met. The first one is ensuring that the field-of-view is consistent with the one used during model creation. The second one is ensuring that the volume is aligned to a common reference system. Notwithstanding further efforts should be devoted to validate it. In this work, we assumed that the inputs images are rigidly aligned. However, in practice, this can not be ensured. Hence, an extension to the current method would be to add a block to verify the orientations and apply a rigid alignment if needed.

With regards to the imaging datasets, both models were created using a 4D MRI dataset acquired from 25 healthy volunteers and validated using an independent hold-out set acquired on 11 cancer patients. This last dataset contained images acquired with a totally different protocol, which allowed to characterize the capacity of the model to work with different image contrasts and appearances. Additionally, other variants were trained and validated using ultrasound images. It is worth mentioning that errors stemming from the reconstruction process, for example discontinuous organ edges between consecutive slices, may negatively affect the image registration process. Therefore, the quality of the employed dataset must be ensured to yield plausible motion fields. Likewise, it influences the accuracy of the deformable registration, which provides the training data for model creation.

While traditional approaches rely on statistical modelling, the proposed models follow a new paradigm that do not require supervised information such as organ segmentation or definition of anatomical landmarks. Establishing inter-subject correspondences is a required step towards the construction of population models. This step is aimed at defining mechanical correspondent landmarks across subjects. Some works have explored different alternatives such as shape-based and landmark-based approaches [217]. However, regardless their nature, this step is time-consuming and prone to errors. Thus, although they have shown promising results, these limitations jeopardize the dissemination of population models. Our methods relies on the strong generalization capability of deep networks to find patterns across a population dataset. Therefore, it replaces the step equivalent to finding inter-subject correspondences with the unsupervised feature learning process. In our opinion, this represents a significant improvement over related methods.

Additionally, basing the motion modelling task on deep neural networks also eases the personalization capability to a new subject since in this context, it would be equivalent to fine-tuning the model's weights. This feature was explored in the works presented in Chapter 7 and Appendix B. The model personalization lead to a better fit to the patient's anatomy and hence an increased accuracy. Therefore, this step is recommended as long as configuration permits. With the progressive expansion of MR-Linacs in the radiotherapy units and the results shown by fast reconstruction strategies [64, 218], having a patient-specific dataset for personalization before treatment is feasible. The accuracy shown by these models is comparable to state-of-the-art algorithms, which normally report average errors of less than 3 mm. In addition, these models introduce other advantages and have shown their capacity to cope with effects such as organ drift, irregular cycles and shifting of the surrogate slice position.

There are however some major differences between the models. In some cases, the differences are improvements introduced by the probabilistic model over the initial modelling solution. The training protocol designed for the deterministic model comprises 3 sequential steps, which was due to the mechanism used to link the surrogates to the motion model. The first step is focused solely on the motion modelling. The second seeks to learn an embedding corresponding to a single future time from dynamic slices while the third joined them all together. In contrast, the model described in Chapter 7 simplifies the loss function and thereby allows to jointly learn all these tasks in a single training step. Hence, in terms of the training protocol, it eliminates the necessity of 3 different steps during training, which can be time-consuming for large datasets.

The methodology followed for the surrogate association represents another important difference between the models. In the first case, the association between the images and 3D deformations is done by computing a L2 distance. Moreover, feature vectors coming from the motion encoder and the surrogate branch are treated independently, thereby requiring the introduction of an additional term in the loss function to minimize their distances. In the second case, we propose to link both sources as a conditional dependency, which is explicitly modeled by the concatenation of the latent phase representation with a temporal embedding. During our experiments we found this is beneficial for the model generalization to unseen cases. Regarding the surrogates, a further extension could be aimed at relating multimodal images since, in theory, This would require the synchronized acquisition of both datasets, similarly as performed in [206] and [180].

In terms of temporal prediction, the first motion model is limited to a single future time point. Depending on the temporal resolution of the images, and the clinical application, this may not be sufficient. In the second model, we integrate a multi-time predictive mechanism able to produce multiple volumes in one shot. In this regard, the temporal forecasting capability is crucial. In this thesis, efforts were devoted to explore solutions for this task, i.e. for future image generation and visual representation forecasting.

## 10.3 Temporal prediction and tracking

Chapter 8 presents a multi-scale recurrent encoder-decoder model, which leverages a differentiable spatial transformation to implicitly learn the future displacement fields. Thanks to the feature extraction stage, the developed network learns in the domain of pixel intensities and implicitly regress the transformations. Therefore, it simultaneously learns to register the images and subsequently extrapolate the deformation ahead of time. Moreover, their capabilities were validated on three imaging modalities, namely MRI, CT and ultrasound, which makes it very valuable for clinical applications. Furthermore, we tested the model capacity to learn composite motion, namely respiratory, cardiac and even peristaltic motion. Our experiments reveal a performance slightly lower than those obtained in slices where most of the movement was breathing-induced, although it remained quite accurate, showcasing the complexity of predicting motion from various sources.

The new generation of temporal predictors is explored in Chapter 9, where we combine the power of convolutional neural networks to extract visual features with an attention-based structure, i.e. the Transformer model, acting on image representations. The visual representations obtained from the images are employed as tokens for the Transformer. Unlike the original model proposed for machine translation, where the queries are composed by the target language, we address the forecasting task by learning the queries. The generation of queries was supported by prior knowledge obtained from ground-truth images, available during model creation. Experimental results showed that conditioning the queries to the prior information yields improved results, which can be attributed to narrowing the prediction space.

This temporal predictor represents a new paradigm compared to existing models. In contrast to previous works, it avoids auto-regression, which is prone to error propagation. Besides, it does not relies on skip connections, which makes it flexible to be integrated into 4D motion models. Experimental results revealed that the proposed model outperforms recurrent predictors. The introduction of convolution operation within the Transformer is an interesting avenue for future work.

Chapter 9 was also describes how a tracker module can benefit from the prediction performed by the model to refine motion fields within a region of interest. We have shown that geometrical errors decrease when using this module. Although we used a single tracking area, it could be extended for multi-region tracking.

## 10.4 Summary

The proposed motion modelling solutions presented throughout this thesis in Chapters 8, 6, 7 and 9 are neither limited to a specific organ, nor to radiation therapy as a treatment modality. In fact, other applications requiring motion compensation can be considered. Nonetheless, there is one consideration that is common to all the methods developed in this thesis. It is related to the nature of the datasets employed. Except a few exceptions, the methods were developed and evaluated mostly with data acquired from healthy volunteers. For a proof of concept, this remove potential adverse effects of liver diseases on the respiration of the patient. Thus, the evaluation environment is more controlled. Nevertheless, future studies should be focused on assessing whether typical diseases targeted in the context of external radiotherapy, such as liver carcinoma, may affect the methods.

From a global perspective, the innovative methodologies proposed in this thesis interconnect some of the typical components present in image-guided interventions. The automatic method developed in Chapter 5, defines an approach for 4D image formation, which is the basic step to create motion models. In addition, the motion models developed in Chapters 6 and 7, establish new solutions to relate partial observation acquired in the treatment room with volumetric deformations. In addition to the actual target tracking, the volumetric deformations are useful for several side tasks such as dose calculation and replanning. In addition, the methods introduced in Chapters 8 and 9 are aimed at compensating the cumulative system latency caused by image acquisition, target localization, and subsequent beam modulation/tracking result.

# CHAPTER 11 CONCLUSION AND RECOMMENDATIONS

This thesis addresses the general problem of respiratory motion management during image-guided radiation therapy. The literature review presented in Chapter 3 highlighted the challenges of 4D image formation, temporal predictive models as well as current motion modelling techniques. It also revealed the current limitations of the state-of-the-art on each one of these topics. Motivated for some of these limitations, a set of tools were developed in order to: (i) observe temporal dynamics of moving organs, (ii) express the high-dimensional deformations on a latent space and link them with partial observation, and (iii) forecast future visual representations. Specifically, the methodology explained in Chapter 5 proposed a fully automatic self-sorting 4D MR volume construction method. In Chapters 6 and 7, we introduced the first motion models, based exclusively on deep neural networks, that can be used as a population-based model, that can be easily personalized. These unsupervised frameworks can generate 4D volumes given only a reference pre-treatment volume and real-time 2D slices. Finally, in Chapters 8 and 9, new strategies were proposed to forecast visual representations and allows future image generation. The main findings and contributions from these research objectives were discussed in Chapter 10. We expect the developed motion modelling approaches to have an impact on precision radiation delivery. Moreover, they may give physicians the confidence to shrink margins while escalating dose and reducing fractions. Therefore, patients would have better treatment outcomes, reduced toxicity and improved quality of life. The next sections state the contributions of the thesis, current limitations, and main recommendations for future work.

## 11.1 Advancement of knowledge

In this thesis we propose several contributions for the purpose of respiratory motion modelling. The first one, presented in Chapter 5, demonstrates the derivation of a pseudonavigator from cine-acquisitions. Moreover, a weighted graph-based approach is introduced for slice stacking. The described methodology represents an alternative for 4D image construction whenever external navigators are not available during image acquisition. The second contribution lies on the elimination of inter-subject correspondences as a previous step during the construction of population-based models. Also, it shows how deep neural networks can be leveraged for respiratory motion modelling. Specifically, in Chapters 6 and 7 we introduced two novel model architectures towards this goal. The first one proposed to relate surrogate images and volumetric deformations by minimizing their distances within a latent space. This idea represents an innovative concept for motion modelling. Similarly, the second model addresses the motion modelling task from a probabilistic point of view. It exposes how the surrogate images and the reference volume can be integrated as predictive variables within the loss function. In fact, both approaches exploited the reference volume to regularize the volumetric generation. These models integrate a temporal predictor module to forecast visual representations, whose structure may change independently of the modelling backbone. Our experiments in several datasets shed light on the decision-making process of the models. We show that, in the latent space, data points are discriminated according to their respiratory phase. Also, in the probabilistic variant, every time the latent space is sampled we can recover a new plausible deformation. Uncertainty maps can therefore be constructed from several generations. These novel features advance the knowledge compared to traditional methods.

The third contribution, discussed in Chapters 8 and 9, presents temporal predictive mechanisms acting on dynamic images. We showed how the forecasted visual representations are the key to drive motion models. The developed models contain cutting edge structures, which are arranged to solve the complex task of estimating future deformations. In Chapters 9 we describe how the latent representations of the motion model, as well as the predicted motion, can be leveraged into a tracker module to improve target location. This approach constitutes an off-the-beaten-path solution.

### 11.2 Limitations

Notwithstanding the advantages of each individual contribution, there are general limitations which should be mentioned:

- A limitation of the graph-based slice sorting approach is that physiological correctness cannot be ensured even if temporal coherence is ensured. Furthermore, due to the nature of the approach, an accurate reconstruction will depend on acquiring enough combinations of the slices at different respiratory states.

- In the developed motion models, we assume that the surrogate images and the reference volume are rigidly align to a common reference space. However, in the clinical scenario this will depend on the scanner settings. Therefore, a previous step should be integrated to ensure this condition is respected.

- Except a few exceptions, the methods developed in this thesis were assessed mostly with data acquired from healthy volunteers. Therefore, potential adverse effects of liver diseases on the respiration of the patient were not considered. Furthermore, since 4D datasets are rather scarce, the evaluation environment was limited by the available amount of subjects.

### 11.3 Future research

To conclude this work, we present the major recommendations, which give the main research lines to explore in future studies.

**Recommendation 1:** Further efforts should be devoted to assess the accuracy of the developed motion models from a dosimetric point of view, similarly as performed in related works [303, 304].

**Recommendation 2:** Although a preliminary evaluation was conducted on a dataset with 11 patients, future studies should be focused on assessing whether typical diseases targeted in the context of external radiotherapy (e.g. liver carcinoma) may affect the methods. Furthermore, the impact of sudden events should be evaluated, such as coughing, sneezing and other sources of involuntary motion, on the models. Perhaps a quality factor could inform about the motion model response to these events.

**Recommendation 3:** Further validation should be done on the performance of the motion models when the reference volume has different acquisition parameters to the ones used in the training set.

**Recommendation 4:** An interesting future direction would be to combine the 4D motion model with a dose prediction framework for online adaptive radiotherapy. Moreover, segmentation maps could be integrated to the motion model whenever they are available.

# REFERENCES

[1] “Mechanics of breathing,” https://teachmephysiology.com/respiratory-system/ventilation/mechanics-of-breathing/, accessed: 2021-11-10.

[2] M. von Siebenthal *et al.*, “4d mr imaging of respiratory organ motion and its variability,” *Physics in Medicine & Biology*, vol. 52, no. 6, p. 1547, 2007.

[3] S. Sayeh *et al.*, “Respiratory motion tracking for robotic radiosurgery,” in *Treating tumors that move with respiration.* Springer, 2007, pp. 15–29.

[4] C. Paganelli *et al.*, “Mri-guidance for motion management in external beam radiotherapy: current status and future challenges,” *Physics in Medicine & Biology*, vol. 63, no. 22, p. 22TR03, 2018.

[5] C. F. Baumgartner *et al.*, “High-resolution dynamic mr imaging of the thorax for respiratory motion correction of pet using groupwise manifold alignment,” *Medical image analysis*, vol. 18, no. 7, pp. 939–952, 2014.

[6] J. Cai *et al.*, “Four-dimensional magnetic resonance imaging (4d-mri) using image-based respiratory surrogate: a feasibility study,” *Medical physics*, vol. 38, no. 12, pp. 6384–6394, 2011.

[7] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[8] E. L. Denton *et al.*, “Unsupervised learning of disentangled representations from video,” in *Advances in neural information processing systems*, 2017, pp. 4414–4423.

[9] Z. Luo *et al.*, “Unsupervised learning of long-term motion dynamics for videos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2203–2212.

[10] L. Castrejon, N. Ballas, and A. Courville, “Improved conditional vrnns for video prediction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7608–7617.

[11] T. Rohlfing *et al.*, “Modeling liver motion and deformation during the respiratory cycle using intensity-based nonrigid registration of gated mr images,” *Medical physics*, vol. 31, no. 3, pp. 427–432, 2004.

[12] A. King *et al.*, “Real-time respiratory motion correction for simultaneous pet-mr using an mr-derived motion model,” in *2011 IEEE Nuclear Science Symposium Conference Record.* IEEE, 2011, pp. 3589–3594.

[13] M. von Siebenthal *et al.*, “Inter-subject modelling of liver deformation during radiation therapy,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2007, pp. 659–666.

[14] T. Nguyen *et al.*, “Adapting population liver motion models for individualized online image-guided therapy,” in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2008, pp. 3945–3948.

[15] World Health Organization. (2021) Cancer. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[16] J. Figueras *et al.*, “Surgical resection of colorectal liver metastases in patients with expanded indications: a single-center experience with 501 patients,” *Diseases of the colon & rectum*, vol. 50, no. 4, pp. 478–488, 2007.

[17] C. A. Arciero and E. R. Sigurdson, “Liver-directed therapies for patients with primary liver cancer and hepatic metastases,” *Current Treatment options in oncology*, vol. 7, no. 5, pp. 399–409, 2006.

[18] A. H. Mahnken, P. L. Pereira, and T. de Baere, “Interventional oncologic approaches to liver metastases,” *Radiology*, vol. 266, no. 2, pp. 407–430, 2013.

[19] D. A. Jaffray and M. K. Gospodarowicz, “Radiation therapy for cancer,” *Disease Control Priorities,*, vol. 3, pp. 239–247, 2015.

[20] M. Baumann, M. Krause, and R. Hill, “Exploring the role of cancer stem cells in radioresistance,” *Nature Reviews Cancer*, vol. 8, no. 7, pp. 545–554, 2008.

[21] J. Mechalakos *et al.*, “Dosimetric effect of respiratory motion in external beam radiotherapy of the lung,” *Radiotherapy and Oncology*, vol. 71, no. 2, pp. 191–200, 2004.

[22] P. J. Keall *et al.*, “The management of respiratory motion in radiation oncology report of aapm task group 76 a,” *Medical physics*, vol. 33, no. 10, pp. 3874–3900, 2006.

[23] K. E. Barret, S. Boitano, and S. M. Barman, *Ganong’s review of medical physiology.* McGraw-Hill Medical, 2012.

[24] F. Preiswerk, "Modelling and reconstructing the respiratory motion of the liver," Ph.D. dissertation, University_of_Basel, 2013.

[25] J. B. West, *Respiratory physiology: the essentials.* Lippincott Williams & Wilkins, 2012.

[26] M. v. Siebenthal, "Analysis and modelling of respiratory liver motion using 4dmri," Ph.D. dissertation, ETH Zurich, 2008.

[27] Y. Seppenwoolde *et al.*, "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*, vol. 53, no. 4, pp. 822–834, 2002.

[28] Y. Fu *et al.*, "Deep learning in medical image registration: a review," *Physics in Medicine & Biology*, vol. 65, no. 20, p. 20TR01, 2020.

[29] J. R. McClelland *et al.*, "A continuous 4d motion model from multiple respiratory cycles for use in lung radiotherapy," *Medical Physics*, vol. 33, no. 9, pp. 3348–3358, 2006.

[30] D. Rueckert *et al.*, "Nonrigid registration using free-form deformations: application to breast mr images," *IEEE transactions on medical imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[31] J. M. Blackall *et al.*, "A statistical model of respiratory motion and deformation of the liver," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2001, pp. 1338–1340.

[32] F. Preiswerk, "Modelling and reconstructing the respiratory motion of the liver," Ph.D. dissertation, University of Basel, 2013.

[33] G. Samei, C. Tanner, and G. Székely, "Predicting liver motion using exemplar models," in *International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging.* Springer, 2012, pp. 147–157.

[34] P. Arnold *et al.*, "3D Organ Motion Prediction for MR-Guided High Intensity Focused Ultrasound." Springer, Berlin, Heidelberg, 2011, pp. 623–630.

[35] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[36] C. Zachiu *et al.*, “An improved optical flow tracking technique for real-time mr-guided beam therapies in moving organs,” *Physics in Medicine & Biology*, vol. 60, no. 23, p. 9003, 2015.

[37] T. Vercauteren *et al.*, “Diffeomorphic demons: Efficient non-parametric image registration,” *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.

[38] K. Brock *et al.*, “Accuracy of finite element model-based multi-organ deformable image registration,” *Medical physics*, vol. 32, no. 6Part1, pp. 1647–1659, 2005.

[39] K. K. Brock *et al.*, “Feasibility of a novel deformable image registration technique to facilitate classification, targeting, and monitoring of tumor and normal tissue,” *International Journal of Radiation Oncology*Biology*Physics*, vol. 64, no. 4, pp. 1245–1254, 2006.

[40] C. Buerger, T. Schaeffter, and A. P. King, “Hierarchical adaptive local affine registration for fast and robust respiratory motion estimation,” *Medical image analysis*, vol. 15, no. 4, pp. 551–564, 2011.

[41] B. D. de Vos *et al.*, “End-to-end unsupervised deformable image registration with a convolutional neural network,” in *Deep learning in medical image analysis and multimodal learning for clinical decision support.* Springer, 2017, pp. 204–212.

[42] G. Balakrishnan *et al.*, “Voxelmorph: a learning framework for deformable medical image registration,” *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[43] A. Sotiras, C. Davatzikos, and N. Paragios, “Deformable medical image registration: A survey,” *IEEE transactions on medical imaging*, vol. 32, no. 7, p. 1153, 2013.

[44] T. Yamamoto *et al.*, “Retrospective analysis of artifacts in four-dimensional ct images of 50 abdominal and thoracic radiotherapy patients,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 72, no. 4, pp. 1250–1258, 2008.

[45] H. C. Urschel *et al.*, *Treating tumors that move with respiration.* Springer, 2007.

[46] Y. Negoro *et al.*, “The effectiveness of an immobilization device in conformal radiotherapy for lung tumor: reduction of respiratory tumor movement and evaluation of the daily setup accuracy,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 50, no. 4, pp. 889–898, 2001.

[47] J. R. McClelland *et al.*, “Respiratory motion models: a review,” *Medical image analysis*, vol. 17, no. 1, pp. 19–42, 2013.

[48] J. Wölfelschneider *et al.*, “Examination of a deformable motion model for respiratory movements and 4d dose calculations using different driving surrogates,” *Medical physics*, vol. 44, no. 6, pp. 2066–2076, 2017.

[49] J. R. McClelland *et al.*, “Respiratory motion models: a review,” *Medical image analysis*, vol. 17, no. 1, pp. 19–42, 2013.

[50] P. Verma *et al.*, “Survey: real-time tumor motion prediction for image-guided radiation treatment,” *Computing in Science & Engineering*, vol. 13, no. 5, pp. 24–35, 2011.

[51] W. Zou, L. Dong, and B.-K. K. Teo, “Current state of image guidance in radiation oncology: implications for ptv margin expansion and adaptive therapy,” in *Seminars in radiation oncology*, vol. 28, no. 3. Elsevier, 2018, pp. 238–247.

[52] C. Hehakaya *et al.*, “Problems and promises of introducing the magnetic resonance imaging linear accelerator into routine care: The case of prostate cancer,” *Frontiers in oncology*, vol. 10, p. 1741, 2020.

[53] S. Mutic and J. F. Dempsey, “The viewray system: magnetic resonance–guided and controlled radiotherapy,” in *Seminars in radiation oncology*, vol. 24, no. 3. Elsevier, 2014, pp. 196–199.

[54] J. Olsen, O. Green, and R. Kashani, “World’s first applicaton of mr-guidance for radiotherapy,” *Missouri medicine*, vol. 112, no. 5, p. 358, 2015.

[55] B. Raaymakers *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*, vol. 62, no. 23, p. L41, 2017.

[56] B. Oborn *et al.*, “Electron contamination modeling and reduction in a 1 t open bore inline mri-linac system,” *Medical physics*, vol. 41, no. 5, p. 051708, 2014.

[57] Y. Tong *et al.*, “Retrospective 4d mr image construction from free-breathing slice acquisitions: A novel graph-based approach,” *Medical image analysis*, vol. 35, pp. 345–359, 2017.

[58] S. Dieterich *et al.*, “Chapter 19 - Respiratory Motion Management for External Beam Radiotherapy,” in *Practical Radiation Oncology Physics*, S. Dieterich

*et al.*, Eds. Philadelphia: Elsevier, 2016, pp. 252–263. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323262095000195

[59] Z. Deng *et al.*, "Four-dimensional mri using three-dimensional radial sampling with respiratory self-gating to characterize temporal phase-resolved respiratory motion in the abdomen," *Magnetic resonance in medicine*, vol. 75, no. 4, pp. 1574–1585, 2016.

[60] J. Tokuda *et al.*, "Adaptive 4d mr imaging using navigator-based respiratory signal for mri-guided therapy," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 59, no. 5, pp. 1051–1061, 2008.

[61] D. Du *et al.*, "High-quality t2-weighted 4-dimensional magnetic resonance imaging for radiation therapy applications," *International Journal of Radiation Oncology* Biology* Physics*, vol. 92, no. 2, pp. 430–437, 2015.

[62] C. K. Glide-Hurst *et al.*, "Four dimensional magnetic resonance imaging optimization and implementation for magnetic resonance imaging simulation," *Practical radiation oncology*, vol. 5, no. 6, pp. 433–442, 2015.

[63] A. Abdelnour *et al.*, "Phase and amplitude binning for 4d-ct imaging," *Physics in Medicine & Biology*, vol. 52, no. 12, p. 3515, 2007.

[64] B. Stemkens, E. S. Paulson, and R. H. Tijssen, "Nuts and bolts of 4d-mri for radiotherapy," *Physics in Medicine & Biology*, vol. 63, no. 21, p. 21TR01, 2018.

[65] G. Remmert *et al.*, "Four-dimensional magnetic resonance imaging for the determination of tumour movement and its evaluation using a dynamic porcine lung phantom," *Physics in Medicine & Biology*, vol. 52, no. 18, p. N401, 2007.

[66] E. Tryggestad *et al.*, "Respiration-based sorting of dynamic mri to derive representative 4d-mri for radiotherapy planning," *Medical physics*, vol. 40, no. 5, p. 051909, 2013.

[67] Y. Liu *et al.*, "T2-weighted four dimensional magnetic resonance imaging with result-driven phase sorting," *Medical physics*, vol. 42, no. 8, pp. 4460–4471, 2015.

[68] T. van de Lindt *et al.*, "A self-sorting coronal 4d-mri method for daily image guidance of liver lesions on an mr-linac," *International Journal of Radiation Oncology • Biology • Physics*, 2018.

[69] M. Georg *et al.*, “Manifold learning for 4d ct reconstruction of the lung,” in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW’08. IEEE Computer Society Conference on*. IEEE, 2008, pp. 1–8.

[70] C. Wachinger *et al.*, “Manifold learning for image-based breathing gating in ultrasound and mri,” *Medical image analysis*, vol. 16, no. 4, pp. 806–818, 2012.

[71] C. F. Baumgartner *et al.*, “Groupwise simultaneous manifold alignment for high-resolution dynamic mr imaging of respiratory motion,” in *International Conference on Information Processing in Medical Imaging*. Springer, 2013, pp. 232–243.

[72] ——, “Self-aligning manifolds for matching disparate medical image datasets,” in *International Conference on Information Processing in Medical Imaging*. Springer, 2015, pp. 363–374.

[73] J. R. Clough *et al.*, “Mri slice stacking using manifold alignment and wave kernel signatures,” in *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. IEEE, 2018, pp. 319–323.

[74] L. V. Romaguera, R. Plantefève, and S. Kadoury, “Quantitative analysis of 4d mr volume reconstruction methods from dynamic slice acquisitions,” in *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 10951. International Society for Optics and Photonics, 2019, p. 1095137.

[75] N. Dikaios *et al.*, “Mri-based motion correction of thoracic pet: initial comparison of acquisition protocols and correction strategies suitable for simultaneous pet/mri systems,” *European radiology*, vol. 22, no. 2, pp. 439–446, 2012.

[76] J. Yang *et al.*, “Four-dimensional magnetic resonance imaging using axial body area as respiratory surrogate: Initial patient results,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 88, no. 4, pp. 907–912, 2014.

[77] Y. Liu *et al.*, “Investigation of sagittal image acquisition for 4d-mri with body area as respiratory surrogate,” *Medical physics*, vol. 41, no. 10, 2014.

[78] C. Paganelli *et al.*, “Liver 4dmri: A retrospective image-based sorting method,” *Medical physics*, vol. 42, no. 8, pp. 4814–4821, 2015.

[79] J. Uh, M. A. Khan, and C. Hua, “Four-dimensional mri using an internal respiratory surrogate derived by dimensionality reduction,” *Physics in Medicine & Biology*, vol. 61, no. 21, p. 7812, 2016.

[80] L. V. Romaguera *et al.*, "Automatic self-gated 4d-mri construction from free-breathing 2d acquisitions applied on liver images," *International journal of computer assisted radiology and surgery*, vol. 14, no. 6, pp. 933–944, 2019.

[81] Y. Tong *et al.*, "Graph-based retrospective 4d image construction from free-breathing mri slice acquisitions," in *Medical Imaging 2014: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 9038. International Society for Optics and Photonics, 2014, p. 90380I.

[82] Y. Hao *et al.*, "Ofx: A method of 4d image construction from free-breathing non-gated mri slice acquisitions of the thorax via optical flux," *Medical Image Analysis*, vol. 72, p. 102088, 2021.

[83] W. Yang *et al.*, "Four-dimensional magnetic resonance imaging with 3-dimensional radial sampling and self-gating–based k-space sorting: early clinical experience on pancreatic cancer patients," *International Journal of Radiation Oncology* Biology* Physics*, vol. 93, no. 5, pp. 1136–1143, 2015.

[84] J. Park *et al.*, "A radial sampling strategy for uniform k-space coverage with retrospective respiratory gating in 3d ultrashort-echo-time lung imaging," *NMR in Biomedicine*, vol. 29, no. 5, pp. 576–587, 2016.

[85] Z. Deng *et al.*, "Improved vessel–tissue contrast and image quality in 3d radial sampling-based 4d-mri," *Journal of applied clinical medical physics*, vol. 18, no. 6, pp. 250–257, 2017.

[86] F. Han *et al.*, "Respiratory motion-resolved, self-gated 4d-mri using rotating cartesian k-space (rock)," *Medical physics*, vol. 44, no. 4, pp. 1359–1368, 2017.

[87] T. Küstner *et al.*, "Self-navigated 4d cartesian imaging of periodic motion in the body trunk using partial k-space compressed sensing," *Magnetic resonance in medicine*, vol. 78, no. 2, pp. 632–644, 2017.

[88] C. Buerger *et al.*, "Nonrigid motion modeling of the liver from 3-d undersampled self-gated golden-radial phase encoded mri," *IEEE transactions on medical imaging*, vol. 31, no. 3, pp. 805–815, 2012.

[89] B. Stemkens *et al.*, "Optimizing 4-dimensional magnetic resonance imaging data sampling for respiratory motion analysis of pancreatic tumors," *International Journal of Radiation Oncology* Biology* Physics*, vol. 91, no. 3, pp. 571–578, 2015.

[90] ——, “Image-driven, model-based 3d abdominal motion estimation for mr-guided radiotherapy,” *Physics in Medicine & Biology*, vol. 61, no. 14, p. 5335, 2016.

[91] ——, “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*, vol. 62, no. 18, p. 7407, 2017.

[92] ——, “A dual-purpose mri acquisition to combine 4d-mri and dynamic contrast-enhanced imaging for abdominal radiotherapy planning,” *Physics in Medicine & Biology*, vol. 64, no. 6, p. 06NT02, 2019.

[93] R. W. Chan *et al.*, “Temporal stability of adaptive 3d radial mri using multidimensional golden means,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 61, no. 2, pp. 354–363, 2009.

[94] C. M. Rank *et al.*, “Respiratory motion compensation for simultaneous pet/mr based on highly undersampled mr data,” *Medical physics*, vol. 43, no. 12, pp. 6234–6245, 2016.

[95] J. N. Freedman *et al.*, “T2-weighted 4d magnetic resonance imaging for application in magnetic resonance–guided radiotherapy treatment planning,” *Investigative radiology*, vol. 52, no. 10, p. 563, 2017.

[96] N. J. Mickevicius and E. S. Paulson, “Investigation of undersampling and reconstruction algorithm dependence on respiratory correlated 4d-mri for online mr-guided radiation therapy,” *Physics in Medicine & Biology*, vol. 62, no. 8, p. 2910, 2017.

[97] L. Feng *et al.*, “Golden-angle radial sparse parallel mri: combination of compressed sensing, parallel imaging, and golden-angle radial sampling for fast and flexible dynamic volumetric mri,” *Magnetic resonance in medicine*, vol. 72, no. 3, pp. 707–717, 2014.

[98] ——, “Xd-grasp: golden-angle radial mri with reconstruction of extra motion-state dimensions using compressed sensing,” *Magnetic resonance in medicine*, vol. 75, no. 2, pp. 775–788, 2016.

[99] L. Feng, N. Tyagi, and R. Otazo, “Mrsigma: Magnetic resonance signature matching for real-time volumetric imaging,” *Magnetic resonance in medicine*, vol. 84, no. 3, pp. 1280–1292, 2020.

[100] N. R. Huttinga *et al.*, “Nonrigid 3d motion estimation at high temporal resolution from prospectively undersampled k-space data using low-rank mr-motus,” *Magnetic resonance in medicine*, vol. 85, no. 4, pp. 2309–2326, 2021.

[101] N. Srivastava, E. Mansimov, and R. Salakhudinov, “Unsupervised learning of video representations using lstms,” in *International conference on machine learning*, 2015, pp. 843–852.

[102] Y. Wang *et al.*, “Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning,” *arXiv preprint arXiv:1804.06300*, 2018.

[103] Y. Li *et al.*, “Flow-grounded spatial-temporal video prediction from still images,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 600–615.

[104] S. Oprea *et al.*, “A review on deep learning techniques for video prediction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[105] C. Vondrick, H. Pirsiavash, and A. Torralba, “Anticipating visual representations from unlabeled video,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 98–106.

[106] K.-H. Zeng *et al.*, “Visual forecasting by imitating dynamics in natural sequences,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2999–3008.

[107] S. Shalev-Shwartz *et al.*, “Long-term planning by short-term prediction,” *arXiv preprint arXiv:1602.01580*, 2016.

[108] P. Luc *et al.*, “Predicting future instance segmentation by forecasting convolutional features,” in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 584–599.

[109] A. Bhattacharyya, M. Fritz, and B. Schiele, “Bayesian prediction of future street scenes using synthetic likelihoods,” *arXiv preprint arXiv:1810.00746*, 2018.

[110] A. Hu *et al.*, “Probabilistic future prediction for video scene understanding,” in *European Conference on Computer Vision.* Springer, 2020, pp. 767–785.

[111] W. Liu *et al.*, “Future frame prediction for anomaly detection–a new baseline,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6536–6545.

[112] S. Xingjian *et al.*, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[113] Z. Liu *et al.*, "Video frame synthesis using deep voxel flow," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4463–4471.

[114] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4194–4202.

[115] M. Ranzato *et al.*, "Video (language) modeling: a baseline for generative models of natural videos," *CoRR*, vol. abs/1412.6604, 2014.

[116] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[117] K. Cho *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[118] H. Wei, X. Yin, and P. Lin, "Novel video prediction for large-scale scene using optical flow," *arXiv preprint arXiv:1805.12243*, 2018.

[119] L. V. Romaguera *et al.*, "Prediction of in-plane organ deformation during free-breathing radiotherapy via discriminative spatial transformer networks," *Medical Image Analysis*, p. 101754, 2020.

[120] J. Walker, A. Gupta, and M. Hebert, "Dense optical flow prediction from a static image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2443–2451.

[121] M. E. Yumer and N. J. Mitra, "Learning semantic deformation flows with 3d convolutional networks," in *European Conference on Computer Vision.* Springer, 2016, pp. 294–311.

[122] N. Watters *et al.*, "Visual Interaction Networks: Learning a Physics Simulator from Video," in *Advances in Neural Information Processing Systems 30.* Curran Associates, Inc., 2017, pp. 4539–4547.

[123] S. Khan *et al.*, "Transformers in vision: A survey," *arXiv preprint arXiv:2101.01169*, 2021.

[124] J. Devlin *et al.*, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[125] A. Radford *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[126] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[127] N. Carion *et al.*, “End-to-end object detection with transformers,” in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[128] Y. Wang *et al.*, “End-to-end video instance segmentation with transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8741–8750.

[129] R. Girdhar *et al.*, “Video action transformer network,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 244–253.

[130] M. Babaeizadeh *et al.*, “Stochastic variational video prediction,” *arXiv preprint arXiv:1710.11252*, 2017.

[131] D. Jayaraman *et al.*, “Time-agnostic prediction: Predicting predictable video frames,” *arXiv preprint arXiv:1808.07784*, 2018.

[132] M. Chaabane *et al.*, “Looking ahead: Anticipating pedestrians crossing with future frames prediction,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2297–2306.

[133] B. Jin *et al.*, “Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4554–4563.

[134] Y. Wang *et al.*, “Probabilistic video prediction from noisy data with a posterior confidence,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 830–10 839.

[135] R. Villegas *et al.*, “Decomposing motion and content for natural video sequence prediction,” *arXiv preprint arXiv:1706.08033*, 2017.

[136] S. Tulyakov *et al.*, “Mocogan: Decomposing motion and content for video generation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1526–1535.

[137] Y. Jang, G. Kim, and Y. Song, "Video prediction with appearance and motion conditions," in *International Conference on Machine Learning.* PMLR, 2018, pp. 2225–2234.

[138] J.-T. Hsieh *et al.*, "Learning to decompose and disentangle representations for video prediction," *arXiv preprint arXiv:1806.04166*, 2018.

[139] V. L. Guen and N. Thome, "Disentangling physical dynamics from unknown factors for unsupervised video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 474–11 484.

[140] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating the future by watching unlabeled video," *arXiv preprint arXiv:1504.08023*, vol. 2, 2015.

[141] J. Yuen and A. Torralba, "A data-driven approach for event prediction," in *European Conference on Computer Vision.* Springer, 2010, pp. 707–720.

[142] J. Walker, A. Gupta, and M. Hebert, "Patch to the future: Unsupervised visual prediction," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 3302–3309.

[143] X. Jin *et al.*, "Video scene parsing with predictive feature learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5580–5588.

[144] E. Denton and R. Fergus, "Stochastic video generation with a learned prior," in *International Conference on Machine Learning.* PMLR, 2018, pp. 1174–1183.

[145] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.

[146] Y. Wu *et al.*, "Future video synthesis with object motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5539–5548.

[147] S. Bengio *et al.*, "Scheduled sampling for sequence prediction with recurrent neural networks," *arXiv preprint arXiv:1506.03099*, 2015.

[148] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[149] I. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[150] N. Kim and J.-W. Kang, "Long-term video generation with evolving residual video frames," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3578–3582.

[151] X. Chen *et al.*, "Long-term video prediction via criticization and retrospection," *IEEE Transactions on Image Processing*, vol. 29, pp. 7090–7103, 2020.

[152] J.-Y. Franceschi *et al.*, "Stochastic latent residual video prediction," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3233–3246.

[153] F. Leibfried, N. Kushman, and K. Hofmann, "A deep learning approach for joint video frame and reward prediction in atari games," *arXiv preprint arXiv:1611.07078*, 2016.

[154] H. Cai *et al.*, "Deep video generation, prediction and completion of human action sequences," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 366–382.

[155] X. Qi *et al.*, "3d motion decomposition for rgbd future dynamic scene synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7673–7682.

[156] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1747–1756.

[157] N. Kalchbrenner *et al.*, "Video pixel networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1771–1779.

[158] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[159] X. Chen *et al.*, "Learning object-centric transformation for video prediction," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1503–1512.

[160] B. Jin *et al.*, "Varnet: Exploring variations for unsupervised video prediction," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 5801–5806.

[161] M. Jaderberg *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[162] X. Liang *et al.*, "Dual motion gan for future-flow embedded video prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1744–1752.

[163] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," *Advances in neural information processing systems*, vol. 29, pp. 64–72, 2016.

[164] X. Jia *et al.*, "Dynamic filter networks," *Advances in neural information processing systems*, vol. 29, pp. 667–675, 2016.

[165] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," *arXiv preprint arXiv:1605.08104*, 2016.

[166] W. Byeon *et al.*, "Contextvp: Fully context-aware video prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 753–769.

[167] A. X. Lee *et al.*, "Stochastic adversarial video prediction," *arXiv preprint arXiv:1804.01523*, 2018.

[168] T. Xue *et al.*, "Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 91–99.

[169] J. M. Blackall *et al.*, "Alignment of sparse freehand 3-d ultrasound with preoperative images of the liver using models of respiratory motion and deformation," *IEEE transactions on medical imaging*, vol. 24, no. 11, pp. 1405–1416, 2005.

[170] Q. Zhang *et al.*, "A patient-specific respiratory model of anatomical motion for radiation treatment planning," *Medical physics*, vol. 34, no. 12, pp. 4772–4781, 2007.

[171] J. Eom *et al.*, "Modeling respiratory motion for cancer radiation therapy based on patient-specific 4dct data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2009, pp. 348–355.

[172] Y. H. Noorda *et al.*, "Subject-specific four-dimensional liver motion modeling based on registration of dynamic mri," *Journal of Medical Imaging*, vol. 3, no. 1, p. 015002, 2016.

[173] W. Harris *et al.*, "A novel method to generate on-board 4d mri using prior 4d mri and on-board kv projections from a conventional linac for target localization in liver sbrt," *Medical physics*, vol. 45, no. 7, pp. 3238–3245, 2018.

[174] J. Pham *et al.*, “Predicting real-time 3d deformation field maps (dfm) based on volumetric cine mri (vc-mri) and artificial neural networks for on-board 4d target tracking: a feasibility study,” *Physics in Medicine & Biology*, vol. 64, no. 16, p. 165016, 2019.

[175] W. Harris *et al.*, “Volumetric cine magnetic resonance imaging (vc-mri) using motion modeling, free-form deformation and multi-slice undersampled 2d cine mri reconstructed with spatio-temporal low-rank decomposition,” *Quantitative Imaging in Medicine and Surgery*, vol. 10, no. 2, p. 432, 2020.

[176] Y. Zhang *et al.*, “A technique for estimating 4d-cbct using prior knowledge and limited-angle projections,” *Medical physics*, vol. 40, no. 12, p. 121701, 2013.

[177] W. Harris *et al.*, “A technique for generating volumetric cine-magnetic resonance imaging,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 95, no. 2, pp. 844–853, 2016.

[178] H. J. Fayad *et al.*, “A generic respiratory motion model based on 4d mri imaging and 2d image navigators,” in *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC).* IEEE, 2012, pp. 4058–4061.

[179] N. Garau *et al.*, “A roi-based global motion model established on 4dct and 2d cine-mri data for mri-guidance in radiation therapy,” *Physics in medicine and biology*, 2019.

[180] A. Giger *et al.*, “Respiratory motion modelling using cgans,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2018, pp. 81–88.

[181] M. von Siebenthal, “Analysis and modelling of respiratory liver motion using 4DMRI,” Ph.D. dissertation, ETH Zurich, 2008.

[182] P. Arnold *et al.*, “3d organ motion prediction for mr-guided high intensity focused ultrasound,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2011, pp. 623–630.

[183] C. Tanner, G. Samei, and G. Székely, “Robust exemplar model of respiratory liver motion and individualization using an additional breath-hold image,” in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI).* IEEE, 2015, pp. 1576–1579.

[184] C. Tanner *et al.*, “In vivo validation of spatio-temporal liver motion prediction from motion tracked on mr thermometry images,” *International journal of computer assisted radiology and surgery*, vol. 11, no. 6, pp. 1143–1152, 2016.

[185] K. Brock *et al.*, “Creating a four-dimensional model of the liver using finite element analysis,” *Medical physics*, vol. 29, no. 7, pp. 1403–1405, 2002.

[186] K. K. Brock *et al.*, “Improving image-guided target localization through deformable registration,” *Acta oncologica*, vol. 47, no. 7, pp. 1279–1285, 2008.

[187] T. He *et al.*, “Online 4-d ct estimation for patient-specific respiratory motion based on real-time breathing signals,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2010, pp. 392–399.

[188] J. Ehrhardt *et al.*, “Statistical modeling of 4d respiratory lung motion using diffeomorphic image registration,” *IEEE transactions on medical imaging*, vol. 30, no. 2, pp. 251–265, 2011.

[189] M. Ries *et al.*, “Real-time 3d target tracking in mri guided focused ultrasound ablations in moving tissues,” *Magnetic Resonance in Medicine*, vol. 64, no. 6, pp. 1704–1712, 2010.

[190] L. V. Romaguera *et al.*, “Predictive online 3d target tracking with population-based generative networks for image-guided radiotherapy,” *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–13, 2021.

[191] ——, “Probabilistic 4d predictive model from in-room surrogates using conditional generative networks for image-guided radiotherapy,” *Medical Image Analysis*, p. 102250, 2021.

[192] L. Vázquez Romaguera, T. Mezheritsky, and S. Kadoury, “Personalized respiratory motion model using conditional generative networks for mr-guided radiotherapy,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2021, pp. 238–248.

[193] Y. H. Noorda *et al.*, “Subject-specific four-dimensional liver motion modeling based on registration of dynamic mri,” *Journal of Medical Imaging*, vol. 3, no. 1, p. 015002, 2016.

[194] C. Hui *et al.*, “4d mr imaging using robust internal respiratory signal,” *Physics in Medicine & Biology*, vol. 61, no. 9, p. 3472, 2016.

[195] M. von Siebenthal *et al.*, “4d mr imaging of respiratory organ motion and its variability,” *Physics in Medicine & Biology*, vol. 52, no. 6, p. 1547, 2007.

[196] Z. Deng *et al.*, “Four-dimensional mri using three-dimensional radial sampling with respiratory self-gating to characterize temporal phase-resolved respiratory motion in the abdomen,” *Magnetic resonance in medicine*, vol. 75, no. 4, pp. 1574–1585, 2016.

[197] Y. Liu *et al.*, “Four dimensional magnetic resonance imaging with retrospective k-space reordering: A feasibility study,” *Medical physics*, vol. 42, no. 2, pp. 534–541, 2015.

[198] N. J. Mickevicius and E. S. Paulson, “Investigation of undersampling and reconstruction algorithm dependence on respiratory correlated 4d-mri for online mr-guided radiation therapy,” *Physics in Medicine & Biology*, vol. 62, no. 8, p. 2910, 2017.

[199] Y. Tong *et al.*, “Retrospective 4d mr image construction from free-breathing slice acquisitions: A novel graph-based approach,” *Medical image analysis*, vol. 35, pp. 345–359, 2017.

[200] X. Chen *et al.*, “High-resolution self-gated dynamic abdominal mri using manifold alignment,” *IEEE Transactions on Medical Imaging*, vol. 36, pp. 960–971, 2017.

[201] M. Modat *et al.*, “Fast free-form deformation using graphics processing units,” *Computer methods and programs in biomedicine*, vol. 98, no. 3, pp. 278–284, 2010.

[202] C. Paganelli *et al.*, “Feasibility study on 3d image reconstruction from 2d orthogonal cine-mri for mri-guided radiotherapy,” *Journal of medical imaging and radiation oncology*, vol. 62, no. 3, pp. 389–400, 2018.

[203] J. Ehrhardt *et al.*, *4D motion modeling: Estimation of respiratory motion for radiation therapy*, ser. Biological and Medical Physics, Biomedical Engineering, J. Ehrhardt and C. Lorenz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 25, no. 4. [Online]. Available: http://link.springer.com/10.1007/978-3-642-36441-9

[204] M. Seregni *et al.*, “Out-of-plane motion correction in orthogonal cine-mri registration,” in *Radiotherapy and Oncology*, vol. 123, 2017, pp. S147–S148.

[205] D. Boye *et al.*, “Population based modeling of respiratory lung motion and prediction from partial information,” in *Medical Imaging 2013: Image Processing*, vol. 8669. International Society for Optics and Photonics, 2013, p. 86690U.

[206] F. Preiswerk *et al.*, “Model-guided respiratory organ motion prediction of the liver from 2d ultrasound,” *Medical image analysis*, vol. 18, no. 5, pp. 740–751, 2014.

[207] C. Jud, F. Preiswerk, and P. C. Cattin, “Respiratory motion compensation with topology independent surrogates,” in *Workshop on imaging and computer assistance in radiation therapy*, 2015.

[208] M. Wilms *et al.*, “Subpopulation-based correspondence modelling for improved respiratory motion estimation in the presence of inter-fraction motion variations,” *Physics in Medicine & Biology*, vol. 62, no. 14, p. 5823, 2017.

[209] R. Girdhar *et al.*, “Learning a predictable and generative vector representation for objects,” in *European Conference on Computer Vision.* Springer, 2016, pp. 484–499.

[210] A. Kurenkov *et al.*, “Deformnet: Free-form deformation network for 3d shape reconstruction from a single image,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV).* IEEE, 2018, pp. 858–866.

[211] P. Coupé *et al.*, “Nonlocal means-based speckle filtering for ultrasound images,” *IEEE transactions on image processing*, vol. 18, no. 10, pp. 2221–2229, 2009.

[212] E. Vorontsov *et al.*, “Boosting segmentation with weak supervision from image-to-image translation,” *arXiv preprint arXiv:1904.01636*, vol. 6, 2019.

[213] T. Mezheritsky, L. V. Romaguera, and S. Kadoury, “3d ultrasound generation from partial 2d observations using fully convolutional and spatial transformation networks,” in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI).* IEEE, 2020, pp. 1808–1811.

[214] S. Klein *et al.*, “Elastix: a toolbox for intensity-based medical image registration,” *IEEE transactions on medical imaging*, vol. 29, no. 1, pp. 196–205, 2009.

[215] L. Henke *et al.*, “Phase i trial of stereotactic mr-guided online adaptive radiation therapy (smart) for the treatment of oligometastatic or unresectable primary malignancies of the abdomen,” *Radiotherapy and Oncology*, vol. 126, no. 3, pp. 519–526, 2018.

[216] C. Kurz *et al.*, “Medical physics challenges in clinical mr-guided radiotherapy,” *Radiation Oncology*, vol. 15, pp. 1–16, 2020.

[217] C. Tanner *et al.*, “Influence of inter-subject correspondences on liver motion predictions from population models,” in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI).* IEEE, 2016, pp. 286–289.

[218] T. Küstner *et al.*, "Cinenet: deep learning-based 3d cardiac cine mri reconstruction with multi-coil complex-valued 4d spatio-temporal convolutions," *Scientific reports*, vol. 10, no. 1, pp. 1–13, 2020.

[219] C. Jud, P. C. Cattin, and F. Preiswerk, "Statistical respiratory models for motion estimation," in *Statistical Shape and Deformation Analysis.* Elsevier, 2017, pp. 379–407.

[220] J. J. Lagendijk *et al.*, "Mri/linac integration," *Radiotherapy and Oncology*, vol. 86, no. 1, pp. 25–29, 2008.

[221] B. Raaymakers *et al.*, "Integrating a 1.5 t mri scanner with a 6 mv accelerator: proof of concept," *Physics in Medicine & Biology*, vol. 54, no. 12, p. N229, 2009.

[222] S. Corradini *et al.*, "Mr-guidance in clinical reality: current treatment challenges and future perspectives," *Radiation Oncology*, vol. 14, no. 1, pp. 1–12, 2019.

[223] W. A. Hall *et al.*, "The transformation of radiation oncology using real-time magnetic resonance guidance: A review," *European Journal of Cancer*, vol. 122, pp. 42–52, 2019.

[224] C. Zachiu *et al.*, "A framework for continuous target tracking during mr-guided high intensity focused ultrasound thermal ablations in the abdomen," *Journal of therapeutic ultrasound*, vol. 5, no. 1, p. 27, 2017.

[225] O. Lorton *et al.*, "Self-scanned hifu ablation of moving tissue using real-time hybrid us-mr imaging," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 8, pp. 2182–2191, 2018.

[226] A. Diodato *et al.*, "Motion compensation with skin contact control for high intensity focused ultrasound surgery in moving organs," *Physics in Medicine & Biology*, vol. 63, no. 3, p. 035017, 2018.

[227] J. Seo *et al.*, "Ultrasound image based visual servoing for moving target ablation by high intensity focused ultrasound," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 13, no. 4, p. e1793, 2017.

[228] H. E. Bainbridge *et al.*, "Treating locally advanced lung cancer with a 1.5 t mr-linac–effects of the magnetic field and irradiation geometry on conventionally fractionated and isotoxic dose-escalated radiotherapy," *Radiotherapy and Oncology*, vol. 125, no. 2, pp. 280–285, 2017.

[229] E. H. Tran *et al.*, "Evaluation of mri-derived surrogate signals to model respiratory motion," *Biomedical Physics & Engineering Express*, 2020.

[230] M. Fast *et al.*, "Tumor trailing for liver sbrt on the mr-linac," *International Journal of Radiation Oncology* Biology* Physics*, vol. 103, no. 2, pp. 468–478, 2019.

[231] B. W. Fischer-Valuck *et al.*, "Two-and-a-half-year clinical experience with the world's first magnetic resonance image guided radiation therapy system," *Advances in radiation oncology*, vol. 2, no. 3, pp. 485–493, 2017.

[232] C. Kontaxis *et al.*, "A new methodology for inter-and intrafraction plan adaptation for the mr-linac," *Physics in Medicine & Biology*, vol. 60, no. 19, p. 7485, 2015.

[233] C. Paganelli *et al.*, "Time-resolved volumetric mri in mri-guided radiotherapy: an in silico comparative analysis," *Physics in Medicine & Biology*, vol. 64, no. 18, p. 185013, 2019.

[234] C. F. Baumgartner *et al.*, "Autoadaptive motion modelling for mr-based respiratory motion estimation," *Medical image analysis*, vol. 35, pp. 83–100, 2017.

[235] G. Samei, C. Tanner, and G. Székely, "Predicting liver motion using exemplar models," in *International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging.* Springer, 2012, pp. 147–157.

[236] X. Han, H. Laga, and M. Bennamoun, "Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[237] J. J. Cerrolaza *et al.*, "3d fetal skull reconstruction from 2dus via deep conditional generative networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2018, pp. 383–391.

[238] C. Biffi *et al.*, "3d high-resolution cardiac segmentation reconstruction from 2d views using conditional variational autoencoders," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019).* IEEE, 2019, pp. 1643–1646.

[239] A. H. Abdi *et al.*, "Variational shape completion for virtual planning of jaw reconstructive surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2019, pp. 227–235.

[240] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in neural information processing systems*, 2015, pp. 3483–3491.

[241] E. D. Brandner *et al.*, "Motion management strategies and technical issues associated with stereotactic body radiotherapy of thoracic and upper abdominal tumors: a review from nrg oncology," *Medical physics*, vol. 44, no. 6, pp. 2595–2612, 2017.

[242] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013.

[243] S. Park *et al.*, "The effect of respiratory baseline drift on the real-time tumor tracking accuracy for liver tumors," *International Journal of Radiation Oncology• Biology• Physics*, vol. 96, no. 2, p. E144, 2016.

[244] A. T. Giger *et al.*, "Liver-ultrasound based motion modelling to estimate 4d dose distributions for lung tumours in scanned proton therapy," *Physics in Medicine & Biology*, 2020.

[245] D. H. Thomas *et al.*, "Initial clinical observations of intra-and interfractional motion variation in mr-guided lung sbrt," *The British journal of radiology*, vol. 91, no. xxxx, p. 20170522, 2018.

[246] M. Mueller and P. Keall, "The markerless lung target tracking challenge (match)," https://www.aapm.org/GrandChallenge/MATCH/, 2019, accessed: 2021-04-20.

[247] A. Yaromina, M. Krause, and M. Baumann, "Individualization of cancer treatment from radiotherapy perspective," *Molecular oncology*, vol. 6, no. 2, pp. 211–221, 2012.

[248] C. Paganelli *et al.*, "Quantification of lung tumor rotation with automated landmark extraction using orthogonal cine mri images," *Physics in Medicine & Biology*, vol. 60, no. 18, p. 7165, 2015.

[249] Y. Ge *et al.*, "Toward the development of intrafraction tumor deformation tracking using a dynamic multi-leaf collimator," *Medical physics*, vol. 41, no. 6Part1, p. 061703, 2014.

[250] C. Ozhasoglu *et al.*, "Synchrony–cyberknife respiratory compensation technology," *Medical Dosimetry*, vol. 33, no. 2, pp. 117–123, 2008.

[251] S. Klüter, "Technical design and concept of a 0.35 t mr-linac," *Clinical and Translational Radiation Oncology*, 2019.

[252] M. Seregni *et al.*, “Motion prediction in mri-guided radiotherapy based on interleaved orthogonal cine-mri,” *Physics in Medicine & Biology*, vol. 61, no. 2, p. 872, 2016.

[253] R. Li *et al.*, “On a pca-based lung motion model,” *Physics in Medicine & Biology*, vol. 56, no. 18, p. 6009, 2011.

[254] F. Preiswerk *et al.*, “Model-guided respiratory organ motion prediction of the liver from 2d ultrasound,” *Medical image analysis*, vol. 18, no. 5, pp. 740–751, 2014.

[255] T.-N. Nguyen *et al.*, “Adapting population liver motion models for individualized online image-guided therapy,” in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008, pp. 3945–3948.

[256] B. Fuerst *et al.*, “Patient-specific biomechanical model for the prediction of lung motion from 4-d ct images,” *IEEE transactions on medical imaging*, vol. 34, no. 2, pp. 599–607, 2014.

[257] Y. H. Noorda *et al.*, “Subject-specific liver motion modeling in mri: a feasibility study on spatiotemporal prediction,” *Physics in Medicine & Biology*, vol. 62, no. 7, p. 2581, 2017.

[258] S. Park *et al.*, “Simultaneous tumor and surrogate motion tracking with dynamic mri for radiation therapy planning,” *Physics in Medicine & Biology*, vol. 63, no. 2, p. 025015, 2018.

[259] G. Meschini *et al.*, “Evaluation of residual abdominal tumour motion in carbon ion gated treatments through respiratory motion modelling,” *Physica Medica*, vol. 34, pp. 28–37, 2017.

[260] C. Tanner *et al.*, “Decision fusion for temporal prediction of respiratory liver motion,” in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2014, pp. 698–701.

[261] D. Rueckert and J. A. Schnabel, “Model-based and data-driven strategies in medical image computing,” *arXiv preprint arXiv:1909.10391*, 2019.

[262] H. Yao *et al.*, “Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5668–5675.

[263] Y. Luo *et al.*, “Lstm pose machines,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5207–5215.

[264] A. Alahi *et al.*, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[265] M. Ranzato *et al.*, "Video (language) modeling: a baseline for generative models of natural videos," *arXiv preprint arXiv:1412.6604*, 2014.

[266] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances in neural information processing systems*, 2016, pp. 613–621.

[267] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3. IEEE, 2004, pp. 32–36.

[268] N. Elsayed, A. S. Maida, and M. Bayoumi, "Reduced-gate convolutional lstm using predictive coding for spatiotemporal prediction," *arXiv preprint arXiv:1810.07251*, 2018.

[269] J. Krebs *et al.*, "Probabilistic motion modeling from medical image sequences: application to cardiac cine-mri," in *International Workshop on Statistical Atlases and Computational Models of the Heart.* Springer, 2019, pp. 176–185.

[270] C. Qin *et al.*, "Joint learning of motion estimation and segmentation for cardiac mr image sequences," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2018, pp. 472–480.

[271] M.-M. Rohé, M. Sermesant, and X. Pennec, "Low-dimensional representation of cardiac motion using barycentric subspaces: A new group-wise paradigm for estimation, analysis, and reconstruction," *Medical image analysis*, vol. 45, pp. 1–12, 2018.

[272] R. Wang *et al.*, "A feasibility of respiration prediction based on deep bi-lstm for real-time tumor tracking," *IEEE Access*, vol. 6, pp. 51 262–51 268, 2018.

[273] T. P. Teo *et al.*, "Feasibility of predicting tumor motion using online data acquired during treatment and a generalized neural network optimized with offline patient tumor trajectories," *Medical physics*, vol. 45, no. 2, pp. 830–845, 2018.

[274] C. Zhao *et al.*, "Predicting tongue motion in unlabeled ultrasound videos using convolutional lstm neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 5926–5930.

[275] H. Lin *et al.*, “Towards real-time respiratory motion prediction based on long short-term memory neural networks,” *Physics in Medicine & Biology*, vol. 64, no. 8, p. 085010, 2019.

[276] F. Azizmohammadi *et al.*, “Model-free cardiorespiratory motion prediction from x-ray angiography sequence with lstm network,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* IEEE, 2019, pp. 7014–7018.

[277] J. Yun, S. Rathee, and B. Fallone, “A deep-learning based 3d tumor motion prediction algorithm for non-invasive intra-fractional tumor-tracked radiotherapy (niftert) on linac-mr,” *International Journal of Radiation Oncology• Biology• Physics*, vol. 105, no. 1, p. S28, 2019.

[278] N. M. De Groot *et al.*, “Three-dimensional catheter positioning during radiofrequency ablation in patients: First application of a real-time position management system,” *Journal of cardiovascular electrophysiology*, vol. 11, no. 11, pp. 1183–1192, 2000.

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

[280] A. V. Dalca *et al.*, “Unsupervised learning for fast probabilistic diffeomorphic registration,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2018, pp. 729–738.

[281] F. Chollet *et al.*, “Keras,” https://github.com/fchollet/keras, 2015.

[282] M. Abadi *et al.*, “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” *arXiv preprint arXiv:1603.04467*, 2016.

[283] V. D. Luca *et al.*, “The 2014 liver ultrasound tracking benchmark,” *Physics in Medicine and Biology*, vol. 60, no. 14, pp. 5571–5599, jul 2015. [Online]. Available: https://doi.org/10.1088%2F0031-9155%2F60%2F14%2F5571

[284] R. Castillo *et al.*, “A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets,” *Physics in Medicine & Biology*, vol. 54, no. 7, p. 1849, 2009.

[285] E. Castillo *et al.*, “Four-dimensional deformable image registration using trajectory modeling,” *Physics in Medicine & Biology*, vol. 55, no. 1, p. 305, 2009.

[286] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[287] J. Cai *et al.*, "Deep lesion tracker: Monitoring lesions in 4d longitudinal imaging studies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 159–15 169.

[288] T. Mezheritsky *et al.*, "Population-based 3d respiratory motion modelling from convolutional autoencoders for 2d ultrasound-guided radiotherapy," *Medical Image Analysis*, p. 102260, 2021.

[289] R. Tao, E. Gavves, and A. W. Smeulders, "Siamese instance search for tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1420–1429.

[290] B. Li *et al.*, "Evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, 2019, pp. 16–20.

[291] A. Gomariz *et al.*, "Siamese networks with location prior for landmark tracking in liver ultrasound sequences," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019).* IEEE, 2019, pp. 1757–1760.

[292] X. Rafael-Palou *et al.*, "Re-identification and growth detection of pulmonary nodules without image registration using 3d siamese neural networks," *Medical Image Analysis*, vol. 67, p. 101823, 2021.

[293] T. Meinhardt *et al.*, "Trackformer: Multi-object tracking with transformers," *arXiv preprint arXiv:2101.02702*, 2021.

[294] M. Danelljan *et al.*, "Adaptive color attributes for real-time visual tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1090–1097.

[295] L. Wang *et al.*, "Visual tracking with fully convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3119–3127.

[296] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4293–4302.

[297] S. Miao, Z. J. Wang, and R. Liao, "A cnn regression approach for real-time 2d/3d registration," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1352–1363, 2016.

[298] N. Wang *et al.*, "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1571–1580.

[299] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[300] R. Xiong *et al.*, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning.* PMLR, 2020, pp. 10 524–10 533.

[301] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[302] S. Lee *et al.*, "Video prediction recalling long-term motion context via memory alignment learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3054–3063.

[303] G. Alina *et al.*, "Liver-ultrasound based motion modelling to estimate 4d dose distributions for lung tumours in scanned proton therapy," *Physics in Medicine & Biology*, vol. 65, no. 23, p. 235050, 2020.

[304] M. Krieger *et al.*, "Liver-ultrasound-guided lung tumour tracking for scanned proton therapy: a feasibility study," *Physics in Medicine & Biology*, vol. 66, no. 3, p. 035011, 2021.

[305] B. Stemkens, E. S. Paulson, and R. H. N. Tijssen, "Nuts and bolts of 4D-MRI for radiotherapy," *Physics in Medicine & Biology*, vol. 63, no. 21, p. 21TR01, oct 2018. [Online]. Available: http://stacks.iop.org/0031-9155/63/i=21/a=21TR01?key=crossref.0f06eeacb332c4300d92339dcfbebf67

[306] Y. Zhang *et al.*, "Preliminary clinical evaluation of a 4d-cbct estimation technique using prior information and limited-angle projections," *Radiotherapy and Oncology*, vol. 115, no. 1, pp. 22–29, 2015.

[307] N. Ballas *et al.*, "Delving deeper into convolutional networks for learning video representations." in *ICLR (Poster)*, 2016.

[308] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

# APPENDIX A QUANTITATIVE ANALYSIS OF 4D MR VOLUME RECONSTRUCTION METHODS FROM DYNAMIC SLICE ACQUISITIONS

## Introduction

Four-dimensional (4D) imaging is a crucial task in several medical applications where the organ motion and breathing-induced anatomical deformation needs to be monitored. During procedures like external beam radiotherapy, respiratory motion can deviate predefined targets and trajectories determined during the treatment planning. Therefore, it is important to characterize and quantify such motion to avoid damages to the healthy tissue. In this context, motion models offer a mean to estimate spatio-temporal displacements of the organ and correct the target position in real time during an intervention. To construct a motion model, data of the entire organ of interest must be acquired. Unfortunately, imaging volumes over time is not a feasible option since it compromises spatial and temporal resolutions. Actually, there is no available implementation of 4D MRI in none of the commercial imaging equipment [305].

Several approaches have been proposed to generate 4D-MRI datasets. Those can be classified as: multi-slice 2D acquisitions and 3D acquisitions. The latter have been reported more recently and interest on it is rising rapidly. Generally, data are acquired over several respiratory cycles, sufficient to capture the motion pattern. A respiratory surrogate is used to binning the data according its respiratory phase to construct volumes over the time [305]. Such auxiliary signal can be either an external surrogate, internal surrogate or self-gating (also known as self-sorting, self-navigation and self-guidance). Some challenges with external and internal surrogates includes low correlation with the internal organ motion and decreasing temporal resolution, respectively. On the other hand, self-gating methods yield a motion signal using features contained in the captured images thus mitigating these shortcomings.

Slice reordering techniques that do not rely on external or internal surrogate signals can be grouped into two main categories: machine learning and slice feature extraction-based methods. Manifold learning (ML) is a machine learning based technique that has shown to be useful in the analysis of motion in medical images [71]. In the context of slice reordering this powerful tool have been employed to map dynamic slices from different anatomical positions into a low-dimensional space according to their respiratory phases. Some methods used to create such representation include: Isomap, Locally Linear Embedding (LLE) and Laplacian Eigenmaps (LE). Moreover, the manifolds yielded from different medical datasets are all combined within one single globally consistent embedding using Manifold Alignment (MA) techniques. Baumgartner et al. (2013) [71] addressed the alignment of multiple manifolds obtained using LLE by overlapping groups of two. They also proposed a sparsification technique for the Gaussian inter-dataset similarity kernel calculation. Later, the authors extended this work [5] by adding a registration-based inter-dataset kernel, which incorporated knowledge of the approximate relations between adjacent slice positions. In Baumgartner et al. (2015) [72], the authors extended the mathematical formulation of LLE to embed more than two datasets simultaneously. They tackled the similarity kernel choice problem by introducing a random walk-based graph matching technique to obtain such kernel. The advantage of that proposal was the global alignment of the data without prior correspondences nor comparisons between the high-dimensional data. Clough et al. (2018) [73] achieved state-of-the-art performance over the former method by introducing a novel graph based descriptor.

In feature extraction methods, obtaining a respiratory signal from the data is the first step in the sorting process. Some authors have proposed to use changes in the body area to generate such signal as it typically correlates with the breathing motion [6,76]. However, this approach is prone to be affected by space-dependent phase shifts. Uh et al. (2016) [79] described a method to yield a self-navigator signal using dimensionality reduction. Nevertheless, the low-dimensional representation of the images does not always proportionally change with respiratory motion thus affecting the correlation between the surrogate signal and organ motion. Van de Lindt et al. (2018) [68] performed a binning of coronal 2D slices according to their cranio-caudal motion to construct 4D volumes. However, the validation is performed against a navigator signal, thus limiting the temporal resolution of their 2D slice series acquisition and makes the binning process easier. A simple and practical graph-based method was presented by Tong et al. (2017) [199] to reconstruct 4D data from the lungs. They constructed a weighted graph considering the slices as nodes and measuring their space and intensity correspondences to finally find the volumes following the shortest paths.

This work aims at quantify the accuracy of two state-of-the-art slice reordering approaches, namely: (1) slice reordering based on Manifold Alignment (MA) with Wave Kernel Signature (WKS) descriptor and (2) slice reordering based on image feature extraction. We also compare the performance to an improved version of the MA method (3). To this end, we introduce spatial metrics and a new temporal metric to assess the coherence of motion in 4D images. The paper is organized as follows. The three 4D MR image construction methods and the proposed metrics are described in Section 2. Experiments and results are presented in Sections 3. Section 4 summarizes our conclusions.

## Material and methods

### Data acquisition

High-resolution sagittal slices under free breathing were acquired on seven volunteers, who provided their written consent. The acquisitions were carried out on a Siemens Skyra 3T scanner using a 2D T2-weighted true FISP sequence with a pixel spacing of $1.7 \times 1.7$ mm$^2$ and a slice thickness of 3 mm. To cover the whole liver, between 66 and 84 slices were acquired, depending on the liver size. Each slice position was imaged 150 times, which cover approximately 5 respiratory cycles, without any gating method.

### Methods for slice reordering

**Manifold Alignment** Manifold Learning is a nonlinear dimensionality reduction technique which aims to map high-dimensional datasets in a low dimensional space, also known as embedding's. Graph-based approaches build connections between two or more disparate data sets subjected to a common process (e.g. free breathing) by aligning their underlying manifolds into a single globally consistent space. Our experiments were performed using the method presented originally by Clough et al. [73] to reorder dynamic slices from the lungs. It yielded the best results in comparison with other MA methods that used different graph descriptors.

Clough's approach uses Laplacian Eigenmaps, which involves creating a Laplacian graph for each dataset to model the internal correspondences inside it. Each dataset is formed by all temporal images, i.e. 150, acquired in a fixed anatomical position. Then, the inter-datasets correspondences are found through a graph based descriptor, names the Wave Kernel Signature. Two alternatives were used for the high dimensional data: raw pixel intensities of the images and the motion fields of each image relatively to an end exhale image. To obtain the motion fields, consecutive temporal images of one slice were registered using NiftyReg generating local displacement fields (df). For each slice $n$, the first end exhale image was identified, and the global motion field between this image and all the other was computed by adding the local deformation fields. In this way, registration between images with large deformations is avoided, thus improving the accuracy. In the aligned embedding given by MA, each point represents a given slice at a time $t$. In the original method [73], the reconstruction stage is performed as follows: starting from one slice $n$, the points on the aligned manifolds are grouped into volumes based on their $L_2$ distance:

$$\forall t_i, \forall m \neq n, stack(m, t_{m_i})\ to\ the\ volume\ at\ t_i\ if\ t_{m_i} = argmin_t(L_2((n, t_i), (m, t)))$$

However, in cases such as the one presented in this paper, with high density in the embedding (more than 10,000 points) due to the high spatial and temporal resolution, it was found this was insufficient to obtain good reconstructed volumes.

**Image feature extraction** Tong et al. [199] proposed a method which uses inter-slice image similarity measure to compute a weighted graph between slices. The shortest path in the graph is used to reorder coronal and sagittal slices of the lungs. The total weight is composed of three weights: $w_g$ that measures the similarity between two slices, $w_s$ that depends on the coherence in the motion direction for both slices and $w_p$ that relates to the difference in position within the respiratory cycle of the slices. The total weight is given as $w = w_g \cdot w_s \cdot w_p$ and ranges between 0 (best) and 1 (worse). For each time, the shortest paths in the graph is found using the Dijkstra algorithm and a 4D volume is reconstructed. For further details, refer to the original method [199]. In this methodology, the first step was the segmentation of the lungs. Since our dataset was centered on the liver, we calculated the image similarity using the whole image.

**Improved Manifold Alignment** We propose an improvement of the first method which consists in comparing the reference point in the manifold $L(n, t_i)$ with the $k$ nearest points on the slice $m$ and select the one that has the highest inter slice image similarity measure with $(n, t_i)$ to be stacked into the volume at time $t_i$ (see Figure A.1). The value of $k$ can increase as the number acquired respiratory cycle for each slice increase. In our experiments,

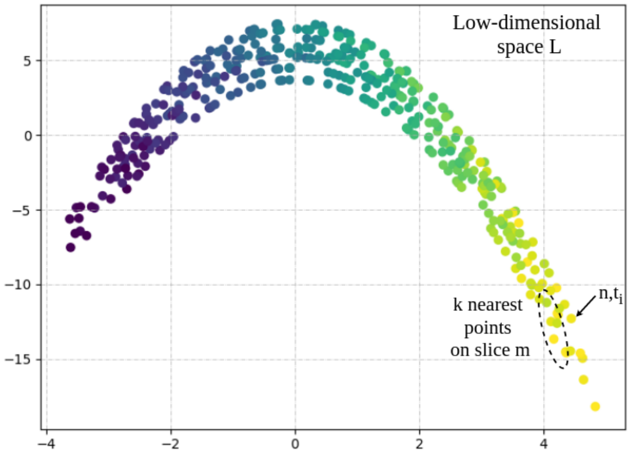


Figure A.1 Aligned manifolds where each point in the low-dimensional space $L$ corresponds to one image at certain anatomical position and time. During the slice stacking the image corresponding to the reference point $L(n, t_i)$ (pointed with the arrow) is compared to the images corresponding to the $k$ nearest points on the slice under analysis $m$.

in which approximately 5 respiratory cycles were acquired for each slice, $k = 5$ proved to be an acceptable value. High values of $k$ will significantly increase computational complexity and impair the discriminative power of the MA.

### Evaluation metrics

To quantify the spatial quality of the volumes, two metrics are proposed. The first is based on the consistency of the diaphragm height across coronal slices, which is measured as the average of the variances calculated in 3 samples from sliding windows:

$$\frac{1}{3(D-1))}\sum_{i=1}^{D-1}\sum_{j=-1}(c_{i+j}-\hat{c}_i)^2 \tag{A.1}$$

where $D$ is the total number of sampled points, $c_k$ represents the height of the diaphragm in the column $k$ and $\hat{c}_k$ the average of $c_{k-1}$, $c_k$, and $c_{k+1}$.

The second metric for spatial quality assessment is an image similarity measure calculated as follow:

$$\sum_{t=0}^{T}\sum_{n=0}^{N}(I_{n,t}-I_{n+1,t})^2 \tag{A.2}$$

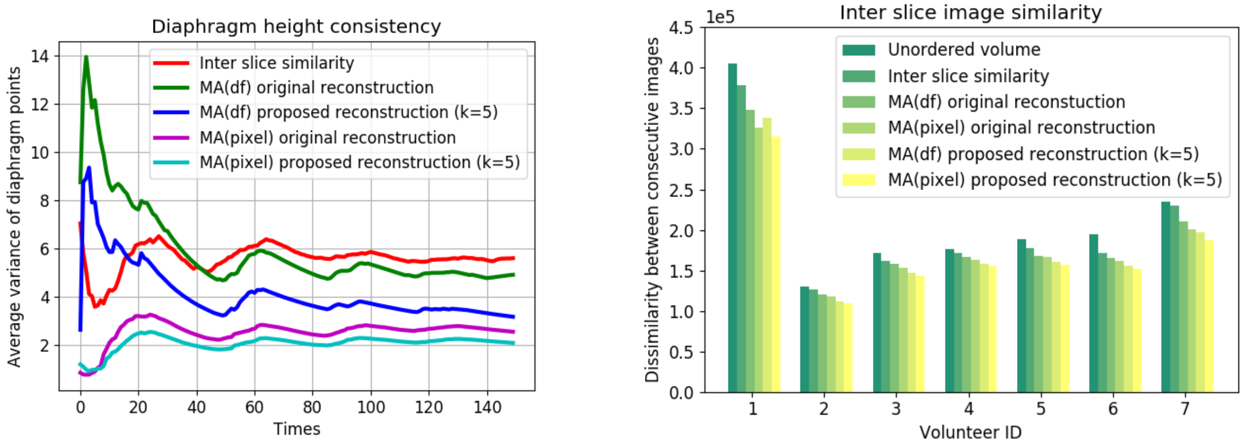
where $I_{n,t}$ is the matrix of pixels intensity in slice $n$ at reconstructed time $t$, $T$ is the number of 3D volumes and $N$ is the number of slices. Note that this metric is a slightly biased toward methods (2) and (3) as a variation of inter slice image similarity measure are also used.

One motion-based metric was implemented to assess the temporal behavior. Since the motion is more evident in the diaphragm area, the standard deviation of the trajectory described by a point at the middle of the right hemi-diaphragm was calculated for each slice. This value was compared before and after reordering.

### Results

Figure A.2 (a) shows the average variance of the diaphragm height in the seven volunteers for each reordering method. The lower values, which were obtained with MA based on pixel intensities combined with our proposed reconstruction, indicate a greater spatial consistency. Also, our method improves the original MA method results regardless to the high dimensionality data used – pixel intensities or motion fields. Interestingly, comparing the manifold-based approaches on with our dataset, the pixel based method outperformed the displacement-based method which reported to yield improved results in Clough et al. [73]. For the second spatial metric, volumes with better reordering presented lower values since the differences between consecutive images are smaller. This shows our method gave the best results as observed in Figure A.2 (b).

Figure A.3 presents the standard deviation of the trajectory followed by the diaphragm. In order to preserve the motion, it is desired that these values be in the same range than the



(a) (b)

Figure A.2 (a) Consistency of the diaphragm points in each 3D volume (b) Total difference between consecutive slices achieved with each method.

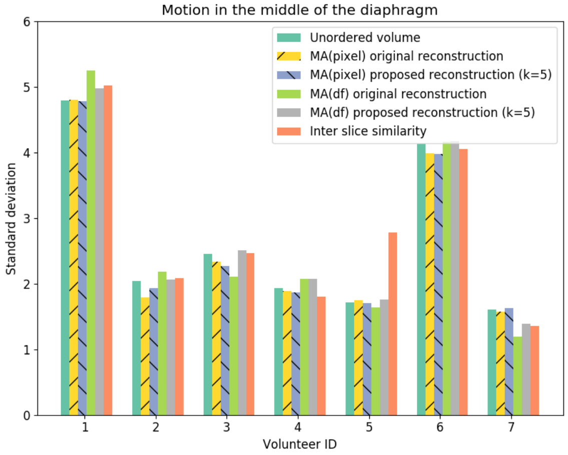


Figure A.3 Measured motion in the middle of the diaphragm for each 4D volume, comparing different reordering methods.

ones shown in the unordered volume. Although generally we can say that the values are similar in all cases, our method showed greater stability in its values.

In general, the best results were obtained with MA methods based on pixel intensities. The reconstruction approach based on closest neighbors proposed in this work proved to improve the quality of the volumes. Figure A.4 shows a qualitative result of the reconstruction achieved in two respiratory states from the stacking of sagittal slices with the proposed method. The reconstruction at inhalation positions is more difficult because the liver does not always descend to the same position.

### Conclusions

In this work, we proposed a new reordering approach and evaluated it against two other approaches using novel tracking-based metrics, one of which is a temporal measure. Results show that our proposed method outperforms state of the art methods in terms of spatial quality and is one of the best, with the MA method based on pixel intensities, in term of temporal quality. The proposed reconstruction scheme from aligned manifolds allows for flexibility in choosing the parameter $k$ depending on the number of points in the embedding. It showed to be useful especially for acquisitions with a high temporal resolution where the discrimination between slices is more challenging. Moreover, it can be applied regardless the imaging modality and the organ to be imaged.

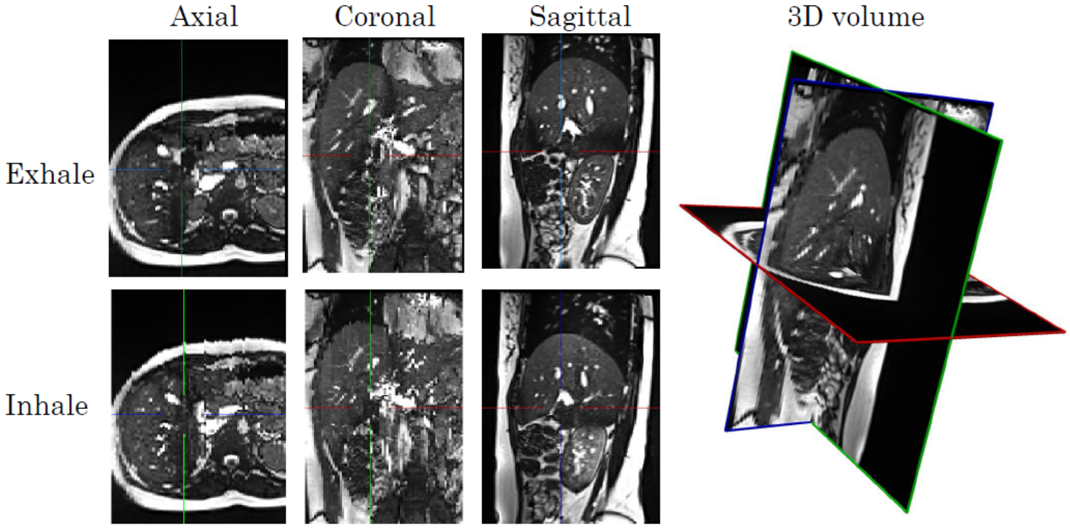


Figure A.4 Volume reconstruction at exhale and inhale respiratory states using pixel based MA with the proposed reconstruction.

# APPENDIX B PERSONALIZED RESPIRATORY MOTION MODEL USING CONDITIONAL GENERATIVE NETWORKS FOR MR-GUIDED RADIOTHERAPY

## Introduction

Shape and motion variability in abdominal and thoracic organs due to the breathing-induced deformation represents an important challenge in external beam radiation therapy. Consequently, tumor tracking and motion compensation strategies are crucial to improve control of radiation beams within the body. Technological innovations such as the MR-Linac have enabled the integration of MR imaging capabilities with linear accelerators into a single device enabling real-time target monitoring during treatment. However, the unobserved out-of-plane motion may degrade dosimetric benefits [215]. Furthermore, volumetric information is useful for dose recalculation and adaptive radiotherapy planning. Current solutions are based on deformable registration [202,204] and statistical motion models [90,174,175,178,179,206,207]. The former strategy applies 2D-3D deformable registration between in-room cine-MRI and a pre-treatment volume to estimate the 3D target position. However, this simple yet effective technique is limited to local motion modeling. Alternatively, several methods are based on maximizing the correlation between a surrogate and a motion model, which can be either population-based or subject-specific. The term surrogate (also known as partial observation) refers to a signal acquired during the intervention, which is directly correlated with the motion of interest [47]. In population-based models, motion data from multiple patients are combined to capture broader motion variability. For instance, Tanner et al. [184] employed a 3D breath-hold scan and interleaved MR slices to drive a statistical model. Although this type of model has shown promising results, their construction involves the challenging task of identifying correspondent landmarks. In contrast, patient-specific models do not require establishing correspondences across a population, providing an improved fit to the patient's anatomy. Typically, the motion extracted from pre-treatment 4D datasets through deformable registration is used to compute a statistical model. Subsequently, partial observations are linked to the model by maximising a similarity metric between the image surrogate and its corresponding slice in the warped reference volume [175, 306]. Often, multi-layer perceptrons are employed to enable ahead-of-time prediction of the model coefficients [174]. Their main limitation is that the weights optimization relies on a pixel-wise similarity metric, which only captures the variation in a single plane [233]. Recent advancements in deep learning have opened new opportunities to relate partial observations to high-dimensional data given sufficiently large training datasets [213, 239, 270]. In the context of motion modelling for image-guided radiation treatments (IGRT), Giger et al. [180] leveraged a conditional generative adversarial network to create a patient-specific model able to relate ultrasound to 3D deformations. However, it lacked interpretability capabilities towards the 3D prediction. Predicting the breathing-induced deformation fields from partial observations has also been explored in 2D [119].

In this work, we propose a predictive framework for abdominal motion, leveraging the advantages of both population-based and patient-specific motion models. During training, the model learns from a population dataset, capturing the wide range in motion variability from multiple anatomies. Moreover, the model's generation capability can potentially benefit from a progressive increase in the amount of data. Once the model is created, it can be personalized to a given patient using relatively few temporal samples, tailoring the model to the subject's specific anatomy at the beginning of the IGRT. In terms of motion modelling, we introduce a novel conditional model which considers the temporal consistency of 2D surrogate images to regress multiple feature representations ahead of time. These feature vectors can be seen as conditioning variables of a low-dimensional space of breathing-induced 3D deformations. Besides, our model has additional advantages compared to related approaches, namely: a latent space capable to discriminate and visualize respiratory phases and the ability to provide uncertainty measures over the model's predictions. Both characteristics make the results more interpretable for clinical procedures.

### Methods

### Model building

During training, our conditional probabilistic model receives as input a single pre-treatment volume gated at a reference respiratory phase and cine-MR images at times $\left\langle t-1,\ t-2,\ \ldots,\ t-m\right\rangle$, which act as predictive variables to recover the dense displacement vector fields (DVF) corresponding to $n$ future respiratory phases. It also receives a set of dense 3D deformations at times $\left\langle t,\ t+2,\ \ldots,\ t+n\right\rangle$. Fig. B.1 shows a schematic representation of the training framework. It is composed of the following blocks: (1) alignment network to generate the DVF, (2) conditional variational autoencoder to learn the motion distribution, and (3) temporal predictor to generate conditioning feature vectors ahead-of-time.

**Alignment network** Since our method does not rely on any surface-based information

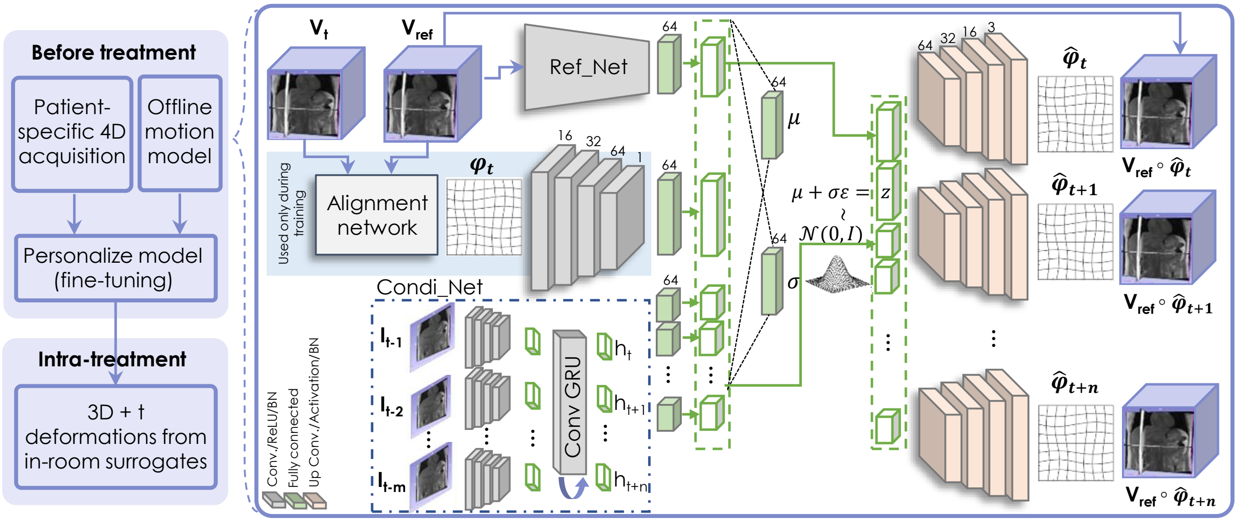


Figure B.1 Proposed motion model used for multi-time volume prediction. The 3D deformations are mapped to a probabilistic latent space, which is conditioned both on extrapolated-in-time vectors and anatomical features. The blue background indicates those components used only during training, whereas the rest are used at all stages.

(i.e. prior segmentations) and avoids explicit voxel generation, we work with deformations between pairs of volumes, from the same subject, over a population dataset. We use a registration function, parameterized with a neural network, which receives a specific reference volume $V_{ref}$ and a target volume $V_t$ at time $t$ as inputs to generate a breathing-induced organ deformation $(\phi_t)$ between them. This deformation is then passed to the following block, which learns the distribution of DVF across the training dataset. In our setup, $V_{ref}$ is taken at the exhale phase since it presents the most reproducible liver representation. We assume that both volumes were previously rigidly aligned to a common reference space. For registration, we use the U-net-like architecture proposed in [42] with pre-trained weights since this step is out of the scope of this work, however other similar configuration can be used.

**Conditional motion field generation** We formulate the 3D volume estimation from partial observations as a conditional manifold learning task, an extension of [240]. The predictive variables, i.e. the pre-operative volume and the cine acquisitions are integrated during optimization in the form of conditional variables which modulate the motion distribution learned by the model. Let $\phi_t$, $V_{ref}$ and $I_s = \left\langle I_{t-1}, I_{t-2}, \ldots I_{t-m} \right\rangle$ be the 3D deformation, the reference volume and the surrogate image sequence, respectively. The goal of the model is to learn the conditional distribution $P\left(\phi_t | I_s, V_{ref}\right)$ to produce a displacement matrix $\hat{\phi}_t \in \mathcal{R}^{H\times W\times D\times 3}$, given the available partial information and subject anatomy, where $H, W$ and $D$ denote the height, width and depth of the volumes, respectively. Following the generative process of conditional variational autoencoders (CVAE), a latent variable $z$ is generated from the prior distribution $p_\theta(z)$ which is constrained to be a Gaussian, i.e. $z \sim \mathcal{N}(0, I)$. By randomly sampling values of $z$, we can generate new DVF. However, computing the posterior distribution $p_\theta(z|I_s, V_{ref})$ to obtain $z$ is analytically intractable [242]. Therefore, an encoder network is adopted to find an approximation of the posterior distribution:

$$q_\psi(z|I_s, V_{ref}) = \mathcal{N}\left(\mu(\phi_t, I_s, V_{ref}), \sigma(\phi_t, I_s, V_{ref})\right). \tag{B.1}$$

This network, parameterized with stacked 3D convolution layers, learns the mean $\mu \in \mathcal{R}^d$ and diagonal covariance matrix $\sigma \in \mathcal{R}^d (d \ll H \times W \times D)$ from the data, as depicted in Fig. B.1. At training, the sampling of $z$ is differentiable with respect to $\mu$ and $\sigma$ by using the "reparameterization trick" [242], and defining $z = \mu + \epsilon * \sigma$, where $\epsilon \sim \mathcal{N}(0, I)$. The distance between both distributions $p_\theta$ and $q_\psi$ can be minimized using the Kullback-Leibler (KL) divergence within a combined loss function which also seeks to minimize a reconstruction loss. The spatial warping block warps the reference volume with the transformation provided by the decoder enabling the model to calculate a similarity measure $\mathcal{L}_{sim}$ between $V_{ref} \circ \phi_t$ and the expected in-room volume $V_t$. We use stochastic gradient descent to find the optimal parameters $\hat{\theta}$ by minimizing the following loss function:

$$\hat{\theta} = \arg\min_\theta \left[\mathcal{L}_{sim}\left(V_{ref} \circ \phi_t, V_t\right) + \mathrm{KL}(q_\psi(z|I_s, V_{ref})||p_\theta(z))\right] \tag{B.2}$$

where the KL-divergence can be computed in closed form. We adopt a negative local cross correlation as similarity loss function. In the proposed architecture, we use a multi-branch convolutional neural network composed by three sub-models that encode: (1) the 3D motion fields provided by the alignment module, (2) the pre-treatment volume ("Ref-Net" sub-network) and (3) the 2D cine image surrogates ("Condi-Net" sub-network). The first and second sub-models possess identical configurations. They are composed of successive 3D convolutions with kernel size 3 × 3 × 3 and a stride of 2 followed, by ReLU activations and batch normalization (BN). On the other hand, Condi-Net acts as temporal predictor. As illustrated in Fig. B.1, each branch ends in a fully connected (FC) layer. The respective outputs are further concatenated and mapped to two additional FC layers to generate $\mu$ and $\sigma$, which are combined with $\epsilon$ to construct the latent space sample $z$, representing the normal Gaussian distribution. The decoder, also modeled with a convolutional neural network, reconstructs the displacement vector fields given the pre-operative volume and the spatiotemporal features extracted from the 2D slices provided in real-time. The conditional dependency is explicitly modeled by the concatenation of $z$ with the feature representation of $V_{ref}$ and $I_s$. This means that our model leverages the transformation retrieved from the latent space given the conditioning feature. Finally, a differentiable layer with explicit spatial transformation capabilities [161] applies the predicted deformation on the pre-operative volume yielding the warped volume that is compared to the target volume in the first term of Eq. (B.2). Using this scheme, our model is able to provide volumetric information.

**Temporal predictor (Condi-Net)** A last module enables multi-time surrogate extrapolation. Its design is inspired by the *seq2seq* configuration, widely applied for natural language processing and other time-series tasks [148]. It is comprised by $m = 3$ stacks of 2D convolutions with kernel size $3 \times 3$ and a stride of 2 followed by ReLU activations and BN. Each stack independently processes the channel-wise concatenation of a single temporal image with their corresponding slice in the pre-operative volume. For a single timestep horizon the result is fed to a convolutional layer. To enable multi-time predictions, the temporal representations are stacked together and fed to convolutional gated recurrent units (GRU) [307], which are arranged in an encoder-decoder configuration. The encoder processes the spatiotemporal features and summarizes the information in a context vector. This embedding is tiled and fed to the decoder, which learns how to extrapolate $n$ feature vectors (depending on the desired number of outputs volume) corresponding to $n$ future time steps.

### Model personalization and application

The goal of the personalization step is to fine-tune the weights of the pre-trained model, which is learned from a population, in order to adapt for the subject-specific anatomy and motion patterns. Thus this process follows a similar methodology as during training, but using a lower initial learning rate on a single subject. During the model application (test stage), the alignment module and the motion encoder are removed. Therefore, the decoder operates as a generative network given the patient anatomy and the cine acquisition, yielding realistic ahead-of-time DVF by sampling $z \sim \mathcal{N}(0, I)$.

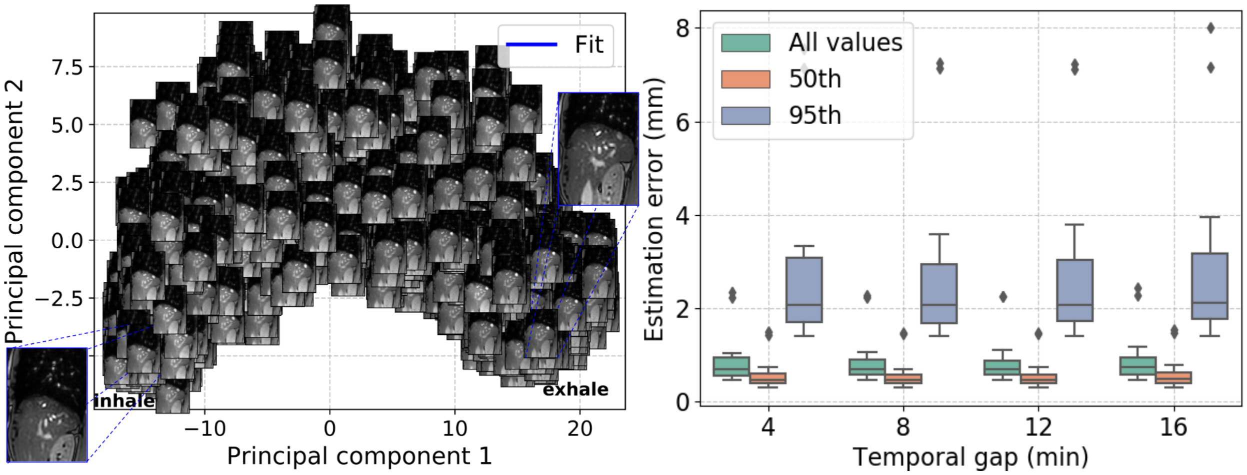
### Experimental Setup and Results

A dataset of free-breathing MR images acquired from a cohort of 25 healthy volunteers, each providing their written consent, was used in this study. Sagittal slices were acquired during 20 min on a MRI clinical scanner (3T Philips Ingenia) using a 2D T2-weighted balanced turbo field echo sequence. Data frames spanning the right liver lobe and navigator slices were acquired following an interleaved scheme and subsequently sorted to create a

Table B.1 Target tracking errors (in mm) measured at different respiratory phases for a predictive horizon of 450 ms. Values are mean ± std. (*Model applied as population-based, i.e. in unseen cases without fine-tuning)

| **Model** | **Mid-inh** | **Inhale** | **Mid-exh** | **Exhale** | **Overall** |
|---|---|---|---|---|---|
| Initial motion | 7.0 ± 6.0 | 7.0 ± 10.34 | 5.3 ± 4.9 | 2.6 ± 2.1 | 5.4 ± 5.8 |
| FM [213] | 3.1 ± 2.6 | 3.9 ± 3.2 | 2.7 ± 2.2 | 1.9 ± 2.1 | 2.9 ± 2.7 |
| ME [202] | 3.0 ± 2.7 | 2.5 ± 2.5 | 2.5 ± 1.7 | 1.7 ± 1.4 | 2.4 ± 2.0 |
| PCA [174] | 1.6 ± 2.0 | 2.0 ± 2.6 | 1.6 ± 0.9 | 2.0 ± 1.2 | 1.8 ± 1.6 |
| Proposed* (C) | 2.4 ± 2.6 | 3.3 ± 3.0 | 2.0 ± 0.9 | 1.3 ± 1.0 | 2.2 ± 1.8 |
| Proposed (S) | 1.8 ± 1.3 | 1.9 ± 1.4 | 1.7 ± 1.2 | 1.5 ± 1.0 | 1.7 ± 1.2 |
| **Proposed (C)** | **1.4 ± 1.0** | **1.8 ± 1.6** | **1.3 ± 1.0** | **1.1 ± 0.8** | **1.4 ± 1.1** |

time-resolved 4D dataset, as detailed in [2]. The in-plane and through-plane resolution was $3.4\times3.4$ mm$^2$ and 3.5 mm, respectively, and image dimension of $32\times64\times64$. For each subject, 80 different sequences of 2D navigators showing different motion amplitudes and frequencies were acquired, which portrays the considerable inter-cycle variability that must be taken into account to increase the robustness of the motion model during radiotherapy. Hence, we leverage this variability as a data augmentation strategy for model creation. The time horizon for a single time step prediction is equivalent to a temporal resolution of 450ms. The



(a) (b)

Figure B.2 (a) Low-dimensional mapping of the latent representations of breathing phases. (b) Analysis of the drift effect on the estimation error when increasing the temporal gap between training and test subsets.

number of volumes for each subject was 2480. We followed a leave-one-out scheme, thereby creating the models with 24 anatomies and the remaining case for personalization/test. Each subject dataset was split into fine-tuning images (5 min, 620 volumes) and test images (15 min, 1860 volumes). We assume that this fine-tuning data represents the treatment planning acquisition, as depicted in the upper left of Fig. B.1. The network's parameters were optimized using the Adam optimizer [308] with an initial learning rate ($lr$) set at $10^{-3}$. For fine-tuning, the $lr = 10^{-5}$ was progressively reduced after 3 epochs without improvements in the validation loss. Training was performed in PyTorch with a batch size of 10.

As a first experiment, between 3 and 5 expert-selected vessel annotations were used to measure the geometrical accuracy between ground-truth (GT) and predicted positions over the last minute ($\approx$ 12 respiratory cycles). These landmark positions were scattered out-of-plane and tracked with subpixel resolution. Two of them were tagged on the same anatomical structure (main portal trunk bifurcation and the first bifurcation of the right portal vein) across all the subjects. The tracking capabilities of the proposed model, with sagittal (S) and coronal (C) orientations used for surrogates, was compared to three state-of-the-art approaches developed in the context of IGRT. Namely, Principal Component Analysis (PCA) [174], a registration-based motion extrapolation (ME) technique [202] and a deep network based on feature merging (FM) [213]. Results presented in Table B.1 were tested for statistical significance using the Wilcoxon signed-rank test with significance level $\alpha = 0.01$. Effect size was measured using Cohen's d. The reference volume was excluded from the error calculation. When comparing the overall tracking errors, the accuracy with the proposed model using coronal images improved by a significant margin of 0.4 mm ($p \ll 0.01$,$d = 0.95$), 1.0 mm ($p \ll 0.01$,$d = 1.02$) and 1.5 mm ($p \ll 0.01$,$d = 0.70$) over PCA, ME and FM, respectively. Moreover, using coronal orientation showed increased performance compared to the sagittal view ($p \ll 0.01$,$d = 0.22$), which is in line with previously reported results [90, 233].

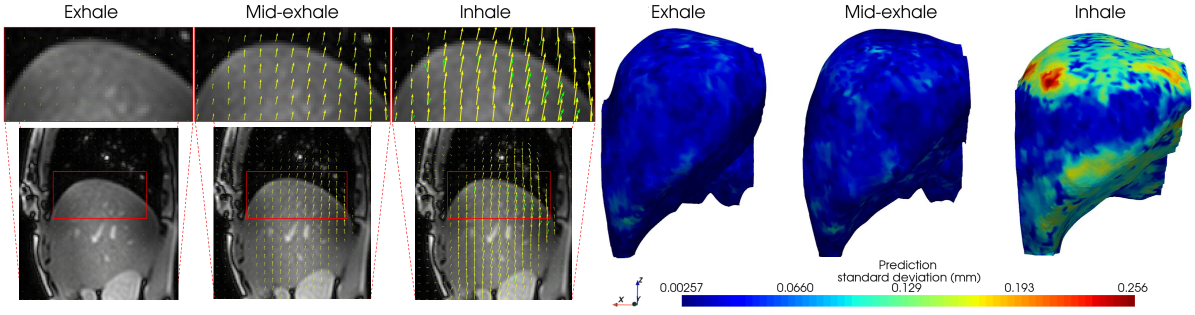


Figure B.3 Left: Most probable deformation (yellow) and reference motion fields (green). Right: Motion-based prediction uncertainty maps at several phases.

We also reported the model's result when applied as population-based, i.e., in unseen cases without prior fine-tuning. Fig. B.2a shows the mapping of the latent vectors to a Cartesian plane via PCA. A clear phase discrimination can be observed, which is plausible for the motion modeling task. When investigating the model's tolerance to potential shifts of the surrogate location, we found little variation of the NCC between GT and predicted volumes with a shift of $\approx$ 20 mm away from the central slice in both directions. This suggests the training process relies primarily on encoded phase information in latent space, an important advantage over current techniques. Each acquisition was divided into 5 equally-sized subsets of 4 minutes in order to investigate the influence of the organ drift. The analysis was conducted by fixing the fifth subset as the testing set, while the first four subsets were used as four separate training sets. The geometrical error distributions reported in Fig. B.2b were measured using 3D deformable registration between ground-truth and predicted volumes with the B-spline transformation model as implemented in the Elastix framework [214]. It can be seen that the error distributions show little to no degradation with the increase of the temporal gap between training and testing sets, showing superior robustness compared to existing techniques where estimation errors are ×4 higher between the extreme subsets [244].

The median [interquartile range] of the error distributions when increasing the horizon to 900 and 1350 ms are 1.8 [2.6] mm and 1.9 [2.9] mm, respectively. The quality of the obtained deformations was assessed through the Jacobian matrix determinant ($|J|$). The percentage of voxels with a non-negative $|J|$ was 99.4% across the entire dataset. Qualitative comparison of the most probable deformation field at selected phases is shown on the left of Fig. B.3,

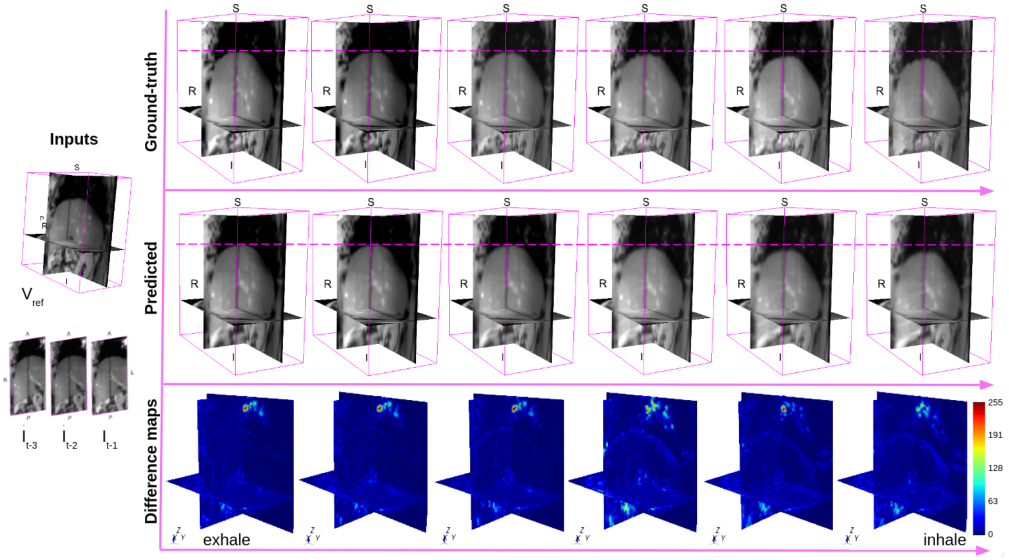


Figure B.4 Difference maps between ground-truth and predicted volumes.

where the reference and the predicted deformations are overlaid. Overall, there is a satisfactory alignment with some minor exceptions. The right part of Fig. B.3 displays uncertainty maps of the predicted DVF, defined as the standard deviation of $N = 50$ different predictions generated by randomly sampling the latent space. Finally, difference maps between GT and predictions across multiple respiratory phases are shown in Fig. B.4. It is noticeable that the model correctly predicts the spatiotemporal motion from inhale to exhale. The proposed method required a mean computation time of 7.44 ms (average from 20 measurements) for inference on a NVIDIA Titan RTX GPU with 64 Gb RAM.

### Conclusion

We presented a novel probabilistic framework for MRI volume predictions during IGRT with a variable predictive horizon from real-time 2D surrogates. It offers several advantages over existing solutions. First, it avoids pre-processing steps such as surface segmentation or landmark annotations. Second, it provides an explainable latent space and quantitative uncertainty metrics, therefore making the results clinically interpretable by physicists. The accuracy of the tracking results are within the clinically acceptable margins (<2mm) for motion management in modalities such as high-intensity focused ultrasound, conventional radiotherapy, or particle therapy. With a prediction horizon of (at least) 450 ms, the motion model is applicable in real-time and meets the typical temporal requirements [203]. Future studies should investigate how the model copes with inter-session variations as well as assessing the dosimetric impact.