UNIVERSITY OF CALIFORNIA

Los Angeles

Deformable Image Registration with Learning

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Bioengineering

by

Yudi Sang

2022

ABSTRACT OF THE DISSERTATION

Deformable Image Registration with Learning

by

Yudi Sang
Doctor of Philosophy in Bioengineering
University of California, Los Angeles, 2022
Professor Dan Ruan, Chair

As a fundamental task in medical image analysis, deformable image registration (DIR) is the process of estimating the deformation vector fields (DVFs) to images. In classic optimization-based DIR method, DVF is solved by optimizing a cost function consisting of image dissimilarity and DVF regularity, which typically involves time-consuming iterative processes. Deep-learning (DL)-based DIR has been developed in recent years, which offers a much faster alternative and the benefit from data-driven regularizing behaviors. This dissertation aims to develop accurate and robust DIR methods and address the lingering challenges in DL-DIR. First, we propose a DIR network that is conscious of and self-adaptive to deformation of various scales to improve accuracy. Second, we propose supervised and unsupervised approaches to incorporate learned implicit feasibility prior into DIR. Third, we propose a domain adaptation method to address the potential domain shift in DIR and improve accuracy and robustness on new data. Finally, we propose a DIR approach to synthesize continuous 4D motion from 3D image pair. Experiments with lung and cardiac images showed that the proposed techniques yielded significant performance improvement. We demonstrate the strength of combining physical-driven rationales and DL techniques in DIR.

The dissertation of Yudi Sang is approved.

Corey W. Arnold

Holden H. Wu

Yingnian Wu

Dan Ruan, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

1 Introduction **1**

2 Scale-Adaptive Deep Network for Deformable Image Registration **6**

2.1 Introduction 6

2.2 Related Works 7

2.2.1 Backbone Architectures of Registraion Networks 7

2.2.2 Dilated Inception 8

2.2.3 Scale Adaptation with Dilation Convolution 9

2.3 Methods 10

2.3.1 The Overall Image Registration Framework 10

2.3.2 Proposed Scale-adaptive Registration Network 11

2.3.3 Dilated Inception Module (DIM) 13

2.3.4 Scale Adaptation Module (SAM) 14

2.4 Experiments and Results 16

2.4.1 Experiment on Cardiac MRIs 17

2.4.2 Experiment on Synthetic Data 21

2.4.3 Experiment on lung 4DCT 24

2.5 Discussion 26

2.6 Conclusion 28

3 Incorporating Feasibility Prior into Deformable Image Registration . . . **30**

3.1 Introduction 30

3.2 Supervised Feasibility Prior Based on Convolutional Auto-encoder 32
3.2.1 Method 32
3.2.2 Experiments on Lung CT and CBCT 36
3.2.3 Experiments on Cardiac CTA and MRI 43
3.3 Unupervised Feasibility Prior Based on Statistical Generative Model 52
3.3.1 Method 52
3.3.2 Experiments on 3D Synthetic Images 59
3.3.3 Experiments on Simulated CT 62
3.3.4 Experiments on 2D Cardiac MRI 65
3.3.5 Experiments on 3D Cardiac MRI 70
3.4 Discussion 73
3.5 Conclusion 76

4 Individualized Test-time Adaptation in Deformable Image Registration 77
4.1 Introduction 77
4.2 Related Works 78
4.2.1 Generalization Gap – Output-space DVF Variations 78
4.2.2 Domain Shift – Input-space Image Variations 79
4.2.3 Test-time Adaptation 80
4.3 Method 82
4.3.1 Baseline Registration Network 83
4.3.2 Test-time Adaptation 83
4.4 Experiments and Results 84
4.4.1 Cross-protocol Adaptation on Lung CBCT 85

4.4.2 Cross-platform Adaptation on Cardiac MRI 88
4.4.3 Cross-modality Adaption on Lung MRI 91
4.5 Discussion 94
4.6 Conclusion 96

5 Continuous 4D Respiratory Motion Synthesis using a Conditional Registration Network 97
5.1 Introduction 97
5.2 Method 98
5.2.1 4D Motion Synthesis 98
5.2.2 Network Architecture 99
5.2.3 Calibration 101
5.3 Experiments and Results 102
5.3.1 Data 102
5.3.2 Implementation 102
5.3.3 Calibration Test 103
5.3.4 Benchmark Methods 103
5.3.5 Evaluation 104
5.3.6 Results 104
5.4 Discussion 107
5.5 Conclusion 110

6 Summary 111

References 114

LIST OF FIGURES

2.1 Overview of the proposed unsupervised registration framework. 10

2.2 Architecture of the scale-adaptive registration network. Numbers inside the blocks indicate number of channels. 12

2.3 Dilated inception module. 13

2.4 Scale adaptation module. "Input from decoding path" is from the lower resolution level, and its path within the module is indicated by the blue arrows. "Input from decoding path" is via the skip connection in the U-net, and its path within the module is indicated by the red arrows. 15

2.5 Example cardiac MRI registration result in our method. (a) Moving image. (b) Fixed image. (c) Warped image. (d) DVF. (e) Visualization of the DRM at the 1/4 level. 18

2.6 Learning curves showing the total loss during the training of the networks. Values for the first two epochs are not displayed. 19

2.7 Histograms showing the distributions of dilation rates in the SAM for testing image pairs with small and large deformations. The number of pixels are the same, and the total counts are equal for the two input settings. 21

2.8 Example synthetic data. (a) Moving image. (b) Fixed image in SimpleElastix. (c) SimpleElastix DVF. (d) Scaled DVF with $\alpha = 0.6$. (e) Forward generated fixed image. 22

2.9 Statistics of the dilation rates for varying magnitudes of deformation introduced by varying α values for data synthesis. 24

2.10 Example thoracic CT registration result in our method. Three coronal slices of a 3D volume is shown. (a) Moving image. (b) Fixed image. (c) Warped image. (d) In-plane components of DVF. (e) Visualization of the DRM at the 1/4 level. 26

3.1 Overview of the proposed CAE-based method. Our method consists of two steps. In the first step, DVFs derived from high-quality images are used to train a feasibility descriptor to capture the underlying feasibility manifold. In the second step, the feasibility descriptor is incorporated into an unsupervised DIR network to regularize the estimated DVFs. 33

3.2 Architecture of the feasibility descriptor. 34

3.3 Architecture of the DVF estimation network. 35

3.4 Example registration results. The first two rows (A) are from a real CBCT, and the last two rows (B) are from a simulated CBCT. 3D DVFs are visualized with their 2D projection onto the coronal plane with the color indicates the deformation magnitude in 3D (unit: mm). White arrows indicate local non-smoothness. Red arrows indicate dubious motions outside the rib cage. 39

3.5 Example input DVF to the feasibility descriptor and its reconstructed output. The background is the corresponding fixed image. Color indicates motion magnitude (unit: mm). The CAE reconstruction effectively removes dubious large motion outside of rib cage (red arrows) and preserves physiological large motion driven by diaphragm dynamics (blue arrows). 40

3.6 Simulated CBCT registration result from networks trained with different balancing weights μ for the feasibility violation loss. The horizontal axis is shown in log scale. 42

3.7 Example motion-compensated image enhancement results on real CBCT. Class BP, Demons still exhibit strong streak artifacts; U-net, U-net BP, Coop CAE, and our method show smoother images but our method has the sharpest detail. 43

3.8 Example motion-compensated image enhancement results on simulated CBCT. The streak artifacts are better alleviated in the four deep learning methods. Sharper detailed structures are reconstructed in our method. 44

3.9 Profiles of the CBCT image before and after enhancement. The image intensity curves are from a horizontal line segment indicated on the image. The numbers indicate the root-mean-square errors to the ground truth for pixels on this line. . 45

3.10 Example MRI image pair and the corresponding segmentation on an axial slice. 46

3.11 Example CTA image and DVF derived from classic B-spline registration. 47

3.12 Example CAE input DVF and its reconstructed output. The background is the corresponding fixed image. 49

3.13 Example registration results on ViewRay data. 50

3.14 Example registration results on cMAC data. The red circle indicates unrealistic motion tangential to the ventricle. 51

3.15 An example case where our method failed to estimate the motion correctly due to the strong artifact (red circle). 52

3.16 Overview of the proposed generator-based registration framework. The trained generator network takes latent deformation variables z as input and outputs a DVF. Then the spatial transformation module warps the moving image towards the fixed with this DVF. During registration, image similarity is calculated, which drives the gradient descent with respect to z to achieve the optimal DVF for each image pair. 55

3.17 Architecture of the generator network (2D version). Numbers above and below the blocks indicate tensor sizes and numbers of channels, respectively. 56

3.18 Training of the generative model. The generator network is trained in an unsupervised setting by alternatively updating z through the inferential back-propagation and θ through the learning back-propagation to maximize the likelihood on the training set $\{(x_i, y_i), i = 1, ..., N\}$. 57

3.19 Example registration results on synthetic images. One slice from a 3D volume is shown. The bottom row shows DVF profiles for the location indicated by the red dashed line. 59

3.20 Boxplots showing the synthetic image registration results. 61

3.21 Example registration results on simulated CT images. Numbers for Elastix registrations indicate the weight λ for bending energy. Red arrows indicate local texture in the subdiaphragmatic region. 62

3.22 Boxplots showing the simulated CT registration results. (a) Target registration error on dense DVFs. (b) Target registration error on anatomical landmarks. . . 65

3.23 Results of the simulated CT registration with different latent variables lengths. (a) Final training loss. (b) Target registration error. 66

3.24 Example registration results on 2D cardiac MRI. Blue dashed arrows indicate the non-smoothness of the DVF. Red solid arrows indicate tissue movements in the myocardium region between the two chambers. 67

3.25 Boxplots showing the results of the 2D cardiac MRI registration evaluated with left ventricle segmentations. (a) Dice coefficient. (b) Average surface distance. (c) Difference between foreground and background deformation magnitude. . . . 69

3.26 Results of the latent varibales interpolation experiment. 70

3.27 Learning curve showing the training loss for the 3D cardiac MRI registration. Only the first 300 epochs are shown. 71

3.28 Example registration results on 3D cardiac MRI using our method. (a) Moving image. (b) Fixed image. (c) One slice from a 3D warped image. The DVF is illustrated with a mesh grid. 72

3.29 Boxplots showing the results of the 3D cardiac MRI registration evaluated with left ventricle segmentations. (a) Dice coefficient. (b) Average surface distance. (c) Difference between foreground and background deformation magnitude. . . . 73

4.1 Overview of the proposed method. The feasibility descriptor is pre-trained and used as a regularizer during the subsequent registration network training. The registration network is initially trained on a set of image pairs, and then refined and adapted to individual image pair in another domain at test time. 82

4.2 Loss curves for the cross-protocol adaptation on lung CBCT. 87

4.3 Example registration results on simulated lung CBCT. 88

4.4 Example motion-compensated image enhancement results on simulated CBCT. The streak artifacts are better alleviated in the network results. Sharper detailed structures are reconstructed after adaptation. 89

4.5 Loss curves for the cross-platform adaptation on cardiac MRI. 91

4.6 Example registration results on cardiac MRI. 92

4.7 An example registration result on cardiac MRI where our method failed to correct the DVF. 92

4.8 Loss curves for the cross-modality adaptation on lung MRI. 94

4.9 Example registration results on lung MRI. 95

5.1 Overview of the proposed motion synthesis method. The conditional registration network takes the two extreme phases of a scan as input and outputs a DVF that correspond to the phase determined by the conditional variable t. The spatial transformer warps the end-exhale image with the DVF. Image dissimilarity to the phase-specific ground truth and DVF regularity are used to drive the network update. 98

5.2 Architecture of the conditional registration network. Conditional residual blocks are used to replace the convolutions in the U-net encoding path and take a scalar parameter t as conditional variable input. 100

5.3 Architecture of the proposed conditional residual block. t is the input conditional variable. a and b are two scalar variables generated from the fully-connected subnetwork. 101

5.4 Example back-search results. The optimal t value has deviated from the "ideal" line, and the shift appears to be global. 104

5.5 Example synthesis results for phase 30%, using the benchmark methods and the proposed approach. Red arrows indicate severe underestimation of motion magnitude. Blue arrows indicate significant local non-smoothness. 107

5.6 Example synthesis results. The numbers indicate the t values. 108

5.7 Curves showing the change of relative motion magnitude along different axes. The average within a region located at the lower lobes was used, and the magnitude was normalized with respect to directional maximum for display. The normalization factors were 4.1, 6.8, and 2.9 mm for the AP, SI, LR respectively. 109

LIST OF TABLES

2.1 Results of the cardiac MRI registration experiment. Results are provided as mean ± standard deviation (Z-value from Wilcoxon signed-rank test). Z-values that indicate statistical significance are underlined. 20

2.2 Results of the synthetic image registration experiment. Results are provided as mean ± standard deviation and p-value from paired t-test. 23

2.3 Results of the lung CT registration experiment. TREs are provided as mean ± standard deviation in millimeter. 27

3.1 Target registration errors based on the anatomical landmarks. Results are provided as mean ± standard deviation in millimeter (p-value from paired t-tests). 41

3.2 Motion-compensated CBCT enhancement results. Results are provided as mean ± standard deviation (p-value from paired t-tests). 46

3.3 Assessment of agreement between structure delineation between warped and fixed images. Results are provided as mean ± standard deviation. Numbers inside the parentheses indicate p-value results from Wilcoxon signed-rank tests when comparing to CAE_{All}. p-values that indicate statistical significance (< 0.01) are underlined. Values in bold indicate the best results. 53

3.4 Target registration errors based on landmarks. Results are provided as mean ± standard deviation in millimeter. Numbers inside the parentheses indicate p-value results from paired t-tests when comparing to CAE_{All}. p-values that indicate statistical significance (< 0.01) are underlined. Values in bold indicate the best results. 54

3.5 Results of the synthetic image registration experiment. Registration errors are provided as mean ± standard deviation and the p-value from paired t-test against the result in our method. 61

3.6 Results of the simulated CT registration experiment. Registration errors are provided as mean ± standard deviation and the p-value from paired t-test against the result in our method. 64

3.7 Results of the 2D cardiac MRI registration experiment. Results are provided as mean ± standard deviation (Z-value from Wilcoxon signed-rank test). Z-values that indicate statistical significance are underlined. 68

3.8 Results of the 3D cardiac MRI registration experiment. Results are provided as mean ± standard deviation (W-value from Wilcoxon signed-rank test). W-values that indicate statistical significance are underlined. 72

4.1 Experimental setup. 84

4.2 Quantitative results on simulated lung CBCT. Results are provided as mean ± standard deviation (p-value from paired t-tests). 88

4.3 Target registration errors based on landmarks on cardiac MRI. Results are provided as mean ± standard deviation in millimeter. Values inside the parentheses indicate p-value results from paired t-tests when comparing to our method. p-values that indicate statistical significance (< 0.01) are underlined. 93

4.4 Target registration errors based on the landmarks on lung MRI. Results are provided as mean ± standard deviation (p-value from paired t-tests). 93

5.1 Impact of the calibration module. Results are provided as mean ± standard deviation (p-value from paired t-tests when comparing the method against ours). 105

5.2 Quantitative results on image synthesis. Results are provided as mean ± standard deviation (p-value from paired t-tests). 106

ACKNOWLEDGMENTS

First, I would like to express my deepest thanks to my advisor, Dr. Dan Ruan, for all her patient guidance throughout the five years. Dr. Ruan has been a perfect mentor, always supporting me with her rich knowledge and insights. She is also caring and understanding, encouraging me to explore as an independent researcher and to grow as a person. I am extremely grateful for having the opportunity to work with her. I would also like to thank my committee members, Dr. Yingnian Wu, Dr. Holden Wu, and Dr. Corey Arnold, for their inspirational discussions and valuable advice.

In addition, I am indebted to my family. I feel sorry for not being able to be at my grandfather's bedside in his last days. I am grateful to my parents, who have always been encouraging and supportive of my journey to pursue the degree. I thank them for believing in my ability to succeed, nurturing me with care and freedom, and accepting me for being myself. I really could not ask for more.

Finally, I would like to thank my dearest friends Yanhan Yao, Wuzhou Tian, and Xinmiao Qu for sharing all the laughter and tears and being there at every moment of my growth. Their unconditional love and support chased away the apprehension I used to have and made me feel that wherever I fly and fall, I can always land safely on the earth of love. Special thanks go to my friend Shua for his company and inspiration during the last year of my PhD study. Shua's passion, pure-hearted self, and unique idealism not only brought me joy but also gently reminded me of what is precious in life and who I truly want to be.

This study was financially supported in part by Varian Medical Systems, Inc. and UCLA Dissertation Year Fellowship.

VITA

2017	B.Eng. (Biomedical Engineering), Beihang University, China
2019	M.S. (Bioengineering), UCLA, Los Angeles, California.
2019–present	Graduate Student Researcher, Department of Radiation Oncology, UCLA.
2021	Teaching Assistant, Department of Molecular, Cell, and Developmental Biology, UCLA. Taught lab sections of MCDB/CSB M130 (Fundamentals of Digital Imaging and Image Processing, Professor Pavak Shah).

PUBLICATIONS

Sang Y and Ruan D. A conditional registration network for continuous 4D respiratory motion synthesis. Med. Phys. (under review).

Sang Y, Cao M, McNitt-Gray M, Gao Y, Hu P, Yan R, Yang Y, and Ruan D. Improving generalization robustness for deep learning-based image registration with individualized test-time adaptation. Med. Phys. (under review).

Fan W, Sang Y, Zhou H, Xiao J, Fan Z, and Ruan D. MRA-free intracranial vessel localization for vessel wall imaging. Sci. Rep. 12.1 (2022): 1-10.

Sang Y, Cao M, McNitt-Gray M, Gao Y, Hu P, Yan R, Yang Y, and Ruan D. Inter-phase 4D cardiac MRI registration with a motion prior derived from CTA. IEEE. Trans. Biomed. Eng. 69.6 (2022): 1828-1836.

Sang Y and Ruan D. Deformable image registration with a scale-adaptive convolutional neural network. Med. Phys. 48.7 (2021): 3815–3826.

Sang Y, Xing X, Wu Y, and Ruan D. Imposing implicit feasibility constraints on deformable image registration using a statistical generative model. J. Med. Imaging 7.6 (2020): 064005.

Sang Y and Ruan D. Synthesizing continuous 4D respiratory motion using a conditional registration network. AAPM, Washington, DC, 2022.

Sang Y, Cao M, McNitt-Gray M, Gao Y, Hu P, Yan R, Yang Y, and Ruan D. Test-time adaptation for deformable image registration. AAPM, Washington, DC, 2022.

Sang Y and Ruan D. 4D-CBCT registration with a FBCT-derived plug-and-play feasibility regularizer. MICCAI, 2021. (Society Travel Award)

Sang Y and Ruan D. Thoracic 4D cone-beam CT registration incorporating a fan-beam CT-derived feasible motion descriptor. AAPM, 2021. (Science Council Session Winner)

Sang Y and Ruan D. Scale-adaptive convolutional neural network for deformable image registration of lung 4DCT. AAPM, 2021.

Fan W, Sang Y, Zhou H, Hu Z, Xiao J, Fan Z, and Ruan D. Angiogram-free intracranial vessel localization for plaque assessment. AAPM, 2021.

Sang Y, Cao M, McNitt-Gray M, Gao Y, Hu P, Yan R, Yang Y, and Ruan D. Enhancing 4D cardiac MRI registration network with a motion prior learned from CTA. ISBI, 2021.

Sang Y and Ruan D. Deformable image registration with a scale-adaptive convolutional neural network. BIBE, 2020.

Sang Y, Xing X, Wu Y, and Ruan D. Deformable image registration using a feasibility-constrained parametrization learned with a statistical generative model. AAPM, 2020.

Sang Y and Ruan D. Enhanced image registration with a network paradigm and incorporation of a deformation representation model. ISBI, 2020.

Sang Y, Xing X, Wu Y, and Ruan D. Imposing implicit feasibility constraints on deformable image registration using a statistical generative model. SPIE Medical Imaging, Houston, 2020.

Sang Y and Ruan D. Imposing learnt flexible vector field prior using a deformation representation model in an image registration network. AAPM, San Antonio, 2019.

CHAPTER 1

Introduction

Physical and physiological motion presents ubiquitously in biomedicine and affects various clinical tasks and applications. Understanding and managing motion has been a fundamental task. Imaging is sensitive to motion. Patient motion during the acquisition can induce artifacts and reduce image quality for diagnosis and analysis. Such motion artifacts manifest as ghosting, blurring, geometric distortion, or decreased signal-to-noise ratio (SNR) [GKS16]. In 4D MRI, motion estimation and compensation techniques have been extensively studied to reorder k-space segments and establish phase correspondence. These techniques typically focuses on "reversing" the motion effect so as to generate a (series of) high-quality snapshot image for radiology reading. On the other hand, motion itself, particularly over a relatively larger field of view (FOV) can provide critical information as well, in tracking spatiotemporal dose deposition in radiotherapy or tumor regression/progression. In such contexts, highly accurate pixel-wise volumetric motion estimation is the central task and demands deformable image registration (DIR).

More precisely, DIR is the process of estimating the deformation vector fields (DVFs) to align two or more images. Typical applications include information fusion across various modalities or setups, motion management, dose accumulation, and longitudinal analysis. The goal of DIR is to achieve accurate point-to-point correspondence. Therefore, its performance is often assessed by target registration errors (TREs) based on dense DVF, anatomical landmarks, or segmentation contours.

DIR has been extensively used in radiation therapy, where understanding tissue place-

ment precisely is critical to control normal tissue toxicity and ensure tumor target coverage. For example, symptomatic (grade $\geq$2) radiation pneumonitis occurs in approximately 30% of patients irradiated for lung cancer, with fatal pneumonitis in about 2% [PST13, JYK12]. Sophisticated beam delivery methods with intensity modulation (i.e., intensity-modulated radiation therapy and volumetric modulated arc therapy) are used to deliver clinically effective dose to the target while minimizing dose to the surrounding normal tissues. Modulated beams generate more conformal dose distribution to a target while sparing nearby normal structures even when those normal structures are located at the concave region of the target [MB06]. However, the conformal dose distribution is only achieved on planning image which is a snapshot of patient anatomy during a long treatment period. Inter- and/or intra-factional anatomy changes from the planning image may deteriorate the dose conformality in actual delivered dose distribution. The use of online imaging, such as cone-beam computed tomography (CBCT) or integrated magnetic resonance imaging (MRI), can detect these anatomical changes and aid in correcting or minimizing the effect of such changes [JSW02, LRR08]. DIR between different phases in a 4D image can provide important information on tissue motion, which help to locate target and normal structures more accurately. In addition, this motion information can be used in function-preserving treatment planning, For example, ventilation information derived from 4DCT DIR can be used in treatment planning to avoid high dose irradiation to the highly functional subregions of the lung.

Classic DIR usually seeks a DVF to minimize a loss function consisting of a dissimilarity measurement between the warped moving image and the fixed image, and regularization energies that penalize undesirable deformations [SDP13, OT14, MBS16].

$$\hat{v} = \underset{v}{\operatorname{argmin}}\{D(F, M \circ v) + R(v)\}, \tag{1.1}$$

where F is the fixed image, M is the moving image, v is the DVF, D is the image dissimilarity, and R is the regularization function. The dissimilarity can be based on image intensity, landmarks, or surface contours [SDP13]. Commonly used regularization functions

encourage prescribed physical behaviors such as smoothness and diffeomorphism to enhance DVF feasibility [SDP13, CDD10, UWS10, HJB10]. Classic DIR usually involves: (1) design or choice of a suitable transformation model and initialization of the associated parameters, (2) use of the transformation model to warp the moving image, (3) evaluation of the image dissimilarity and the regularization function, and (4) update of the parameters in the transformation model by optimizing the cost function, using a suitable optimisation algorithm. One limitation of classic DIR method is that the iterative solving process is often too slow for real-time applications.

Deep learning (DL) approaches have been developed for DIR recently. The registration deep network infers DVFs directly from input image pairs with high efficiency. Training of the network can be supervised or unsupervised. In the supervised setting, the network learns the map between the pair of image input and the corresponding ground-truth DVF directly [YKS17]. In the unsupervised learning methods, a spatial transformation module is employed in the network architecture [JSZ15]. Warped image and image similarity are computed within the training process. Therefore, unsupervised learning-based registration is similar to the classic intensity-based method in the optimization paradigm but replaces the iterative DVF solving process with fast inference from a trained network [BZS19, VBV19].

This dissertation contributes to the field of medical image processing by developing physical-driven rationales which are further integrated with DL techniques. We designed our experiments for inter-phase DIR, aiming to accurately delineate tissue motion from 4D images, but the proposed methods generalizes to other settings. The proposed techniques address the major challenges in DL-DIR from different aspects:

Multi-resolution hierarchical strategy is typically used in classic optimization-based image registration to capture varying magnitudes of deformations while avoiding undesirable local minima. However, without such a strategy, DIR networks are supposed to address all deformation scales with a single reference. A rough concept of the scale is captured in deep networks by the reception field of kernels, and it has been realized to be both desirable

and challenging to capture convolutions of different scales simultaneously in registration networks. In Chapter 2, we propose a DIR network that is conscious of and self-adaptive to deformation of various scales to improve registration performance.

A major challenge for DIR to achieve physically and physiologically sound deformation lies in the mathematical quantification of desirable clinical properties. Despite the long efforts in designing deformation parameterization models and regularization functions for invertibility, volume preservation, diffeomorphism and interface discontinuity preservation, an ideal solution remains elusive. At the center of this challenge is spatial heterogeneity. Specifically, tissue properties including elasticity and sliding directionality, vary across the domain of interest. Classic optimization-based DIR often allows the tuning of a few hyper-parameters to balance the tradeoff between fidelity and the regularization property. Spatial variation in tissue properties means the "optimal" local tradeoff should also be spatially varying, and cannot be fully characterized with only a few hyper-parameters. Therefore, to further improve registration performance, it is necessary to exploit additional prior knowledge in the transformation domain. In Chapter 3, we propose two different approaches to incorporate learned implicit feasibility conditions into DIR.

In classic optimization-based DIR, DVF is solved for each image pair through a time-consuming iterative process. DL-DIR offers a much faster alternative and can benefit from data-driven regularizing behaviors. However, it is possible that training and testing samples differ in either image or motion characteristics or both, resulting in a generalization gap that risks the reliability of direct inference result for each individual test case. Currently most DL-DIR methods impose restrictions on setups, which limits wide model dissemination. Application to new or less common modalities is challenging without sufficiently large training cohort. In Chapter 4, we propose a domain adaptation method to address the potential domain shift, and improve the accuracy and robustness of registration.

4D imaging provides important physiological information for diagnosis and treatment. On the other hand, its acquisition could be challenged with artifacts due to motion or sort-

ing/binning, time and effort bandwidth, and imaging dose considerations. A 4D synthesis development would significantly augment the available data, addressing quality and consistency issues. Furthermore, the high-quality synthesis can serve as an essential backbone to establish a feasible physiological manifold to support online reconstruction, registration, and downstream analysis from real-time imaging. In Chapter 5, we propose a DIR approach to synthesize continuous 4D motion from two extreme phases.

CHAPTER 2

Scale-Adaptive Deep Network for Deformable Image Registration

2.1 Introduction

Despite the benefit of efficiency, it is challenging for DL-DIR methods to achieve highly physically and physiologically feasible registration since deformations of various scales need to be addressed by the network simultaneously. Tissue properties including elasticity and anisotropic discontinuity or sliding, vary across scales and spatial locations. This spatial heterogeneity in deformation scale is challenging for deep networks (CNNs), which typically use a single kernel size with a unique reception field at each level.

One way to address this problem is to implement the multi-resolution strategy by sequentially optimizing multiple networks with the warped image from the previous network passed to the subsequent network as the moving image input to progressively refine the DVF [SCX18, VBV19, ZDC19, CGW19, WAH20]. This increases the total number of network parameters and requires a prolonged training process. In addition, the repeated resampling of the moving image may accumulate errors and thus limit registration accuracy [WAH20].

Another way to better accommodate the scale heterogeneity is by using more sophisticated network architectures. U-net was proposed for segmentation tasks and was shown able to achieve good performance with relatively small datasets [RFB15]. Its hierarchical structure and skip connections between the encoder and decoder allow fusion of information across scales. The U-net structure has also been tested in registration as the backbone in

many studies (details in Sec. 2.2.1). Inception module was proposed with GoogleNet to capture multi-scale contextual features by using kernels with multiple sizes in parallel [SLJ15]. Several versions of the inception module have been proposed in combination with factorized convolutions and residual blocks to improve the performance and efficiency [SVI16, SIV17].

In this study, we propose a network for end-to-end DIR that is conscious of and self-adaptive to information from various scales. Dilated inception modules (DIMs) are proposed to accommodate the need for larger reception fields and a wide range of scales efficiently. Scale adaptation modules (SAMs) are proposed to learn a spatial map of optimal scale based on dilated convolution and combine shallow and deep features in a self-adaptive setting.

2.2 Related Works

2.2.1 Backbone Architectures of Registraion Networks

Pooling-based and U-net-based architectures are two types of backbone networks that are commonly used in both supervised and unsupervised learning DIR methods. Here, we briefly review some studies that addressed the scale problem.

The pooling-based architecture refers to the networks with shrinking feature map width. The networks typically consist of convolution and pooling layers only, and the output DVF size is a fraction of the input size, resulting in a lower degree of freedom. Sokooti et al. [SVB17] proposed a supervised RegNet to predict deformation vectors from input image patches. The network consisted of two paths of convolution and pooling layers to integrate information from two scales. Cao et al. [CYZ18] proposed a cue-aware deep regression network that operates on 3D patches. The network employed a scale-adaptive local similarity as contextual guidance and learned from deformations generated using a sampling strategy in both image and deformation spaces. De Vos et al. [VBV19] used a pooling-based architecture in their unsupervised method to infer B-spline control points from image patches. They stacked multiple networks with different B-spline grid spacing and image resolution at

each level to perform coarse-to-fine DIR.

Compared to the pooling-based approach that outputs sparse DVF interpolated by B-spline or thin-plate spline kernels, U-net [RFB15] is a more popular choice of the backbone network, which predicts dense DVFs with deconvolution or unpooling layers while combining information across scales. With a U-net-based architecture, Rohé et al. [RDH17] proposed SVF-Net to predict velocity fields by learning from segmentation-derived ground-truth deformations in a supervised setting. Balakrishnan et al. [BZS19] proposed VoxelMorph for unsupervised DIR and tested with brain MRIs. They also integrated the method with a diffeomorphic parameterization enabled by the scaling and squaring layers [DBG18]. Zhao et al. [ZLL19] proposed VTN with a cascading scheme, an integration of an affine registration network, and an invertibility loss. They also investigated the recursive cascade of VoxelMorph and VTN and achieved significant improvements with the progressive deformations [ZDC19]. Eppenhof et al. [ELV19] proposed a progressive learning scheme to enable training on large and small deformations with the same supervised U-net. They progressively expanded the network with additional layers on higher resolutions and trained the network using lung CT images with simulated deformations. Cheng et al. [CGW19] proposed to cascade a series dilated convolutions as a refinement network after coarse registration by U-net, to obtain different sizes of receptive fields while maintaining the resolution of feature maps.

2.2.2 Dilated Inception

Dilated inception is the combination of dilated convolutions and inception module. This design has been proposed independently by a few studies to achieve abundant receptive field size without compromising the computational efficiency. The detail of its structure will be introduced in Sec. 2.3.3. Some studies also investigated atrous spatial pyramid pooling (ASPP) and other ways to combine convolutions with different dilation rate and variations to the dilated inception design. Here, we briefly review some studies that use multi-rate dilated convolution layers in computer vision tasks and medical image segmentation tasks.

Shi et al. [SJZ17] proposed a deep network, which cascaded multiple dilated convolution based inception modules, for single image super-resolution. Chen et al. [CPK17] proposed DeepLab for semantic image segmentation with ASPP, which probed an incoming feature map with filters at multiple sampling rates and effective fields-of-views to capture objects and image context at multiple scales. Li et al. [LYG19] proposed a dilated-inception net (DIN) to extract and aggregate multi-scale features for right ventricle segmentation. Yang et al. [YLJ19] proposed a dilated inception module to efficiently capture multi-scale saliency-influential factors for visual saliency prediction. Wang et al. [WLT19] proposed a 3D networks for MRI prostate segmentation with group dilated convolution to aggregates multi-scale contextual information. They explored different ways including sequence, summation, and concatenation, to combine convolutions with different dilation rates. Li et al. [SGK20] introduced dilated inception into the encoder part of U-net to improve liver and tumor segmentation performance. Fu et al. [FLW20b] integrated dilated inception and self-attention gates into the unsupervised learning framework for 4DCT lung DIR and achieved comparable performance to traditional DIR.

2.2.3 Scale Adaptation with Dilation Convolution

Scale adaptation can be achieved by varying the dilation rate or adjusting the combination of dilated convolution kernels, and thereby the size of receptive field. This idea has been investigated in several studies in semantic segmentation and medical image segmentation. Zhang et al. [ZTZ17] proposed scale-adaptive convolutions to obtain flexible-size receptive fields for scene parsing. The scale-adaptive convolutions employed implicitly learned scale coefficient maps to scale the sizes of convolutional patches. Zhang et al. [ZZL20] proposed ASCNet for microscopy image segmentation. A sub-network was used to adaptively learn an appropriate dilation rate for each pixel in the image. The dilation rate was then transmitted as a shared parameter for all convolution layers. Guo et al. [GCY20] proposed a network for automatic melanoma segmentation with knowledge aggregation modules (KAMs) to alleviate

the hole and shrink problems. In KAM, dilated convolutions on shallow features had adaptive receptive fields, which were adjusted according to deep features. Jin et al. [JLZ21] proposed CASINet for scene parsing. In the contextual scale interaction (CSI) module, they used weighted combination of the multi-scale features from ASPP to mimic the use of adaptive filters for each spatial position of each scale. Then the features were further fused using scale and channel attention in the scale adaptation module.

2.3 Methods

2.3.1 The Overall Image Registration Framework

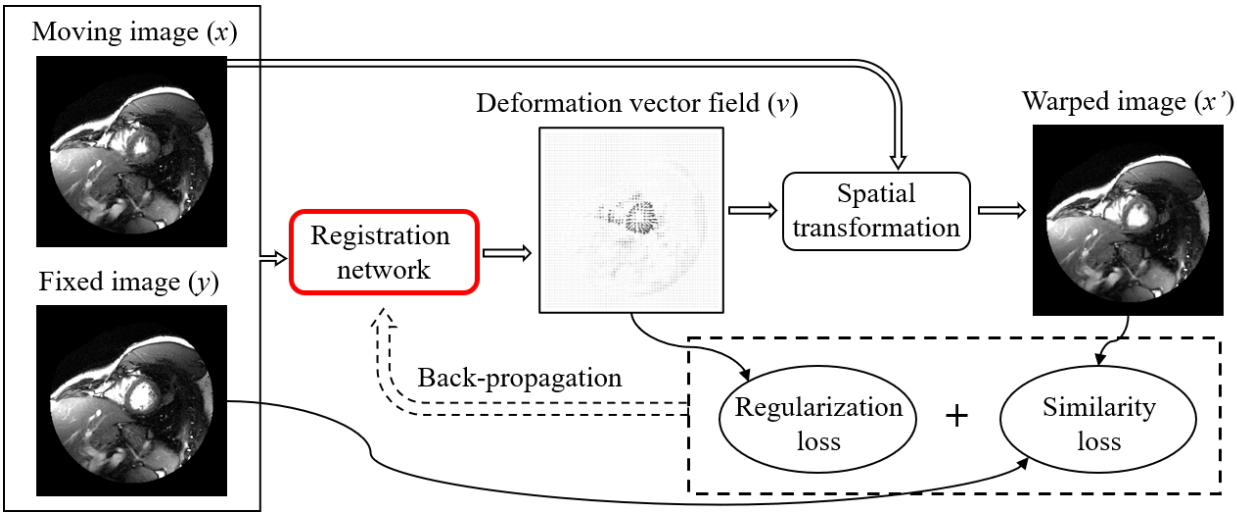

Figure 2.1: Overview of the proposed unsupervised registration framework.

As shown in figure 2.1, the proposed method consists of a registration network and a spatial transformation module. The registration network concatenates fixed and moving images as input and generates a DVF v, with which the moving image x is warped toward the fixed image y. Inside the spatial transformation module, a sampling grid is created using the input DVF. The input moving image x is sampled at these grid points to form the output warped image x' [JSZ15].

The loss function is defined as the weighted sum of the intensity match discrepancy and the regularity penalty:

$$L = L_s(y, x') + \lambda L_r(v), \tag{2.1}$$

where L_s is the image similarity loss, L_r is the DVF regularization loss, and λ is a balancing hyper-parameter. In this work, we use normalized cross-correlation (NCC) as the image similarity metric, but the method also applies to other choices such as mutual information [VVS20].

$$L_s(y, x') = 1 - \mathrm{NCC}(y, x') = 1 - \left\langle \frac{y - \bar{y}}{\|y - \bar{y}\|_2}, \frac{x' - \bar{x'}}{\|x' - \bar{x'}\|_2} \right\rangle. \tag{2.2}$$

Bending energy penalty is used as the DVF regularization loss to penalize nonsmooth deformations and therefore encourage physical feasibility [RSH99]. Its 2D version can be written as:

$$L_r(v) = \frac{1}{d_a d_b} \sum_{a=1}^{d_a} \sum_{b=1}^{d_b} \left[\left(\frac{\partial^2 v}{\partial a^2} \right)^2 + \left(\frac{\partial^2 v}{\partial b^2} \right)^2 + 2 \left(\frac{\partial^2 v}{\partial a \partial b} \right)^2 \right], \tag{2.3}$$

where a and b are spatial indices of the 2D image, and d_a and d_b are the corresponding spatial resolutions, respectively. During training, a back-propagation scheme is used to derive a DVF solution to minimize the objective in equation 2.1.

The 2D and 3D versions of the method are with this common framework. Here in Sec. 2.3, we mainly present the 2D version. The 3D version uses the same number of channels and 2D operations replaced by 3D counterparts.

2.3.2 Proposed Scale-adaptive Registration Network

The major contribution of this work is a scale-adaptive registration module. The module uses a general U-net structure [RFB15] to take advantage of the hierarchical structure and skip connections for effective learning of features at all scales. Scale adaptation is achieved with the introduction of DIMs and SAMs.

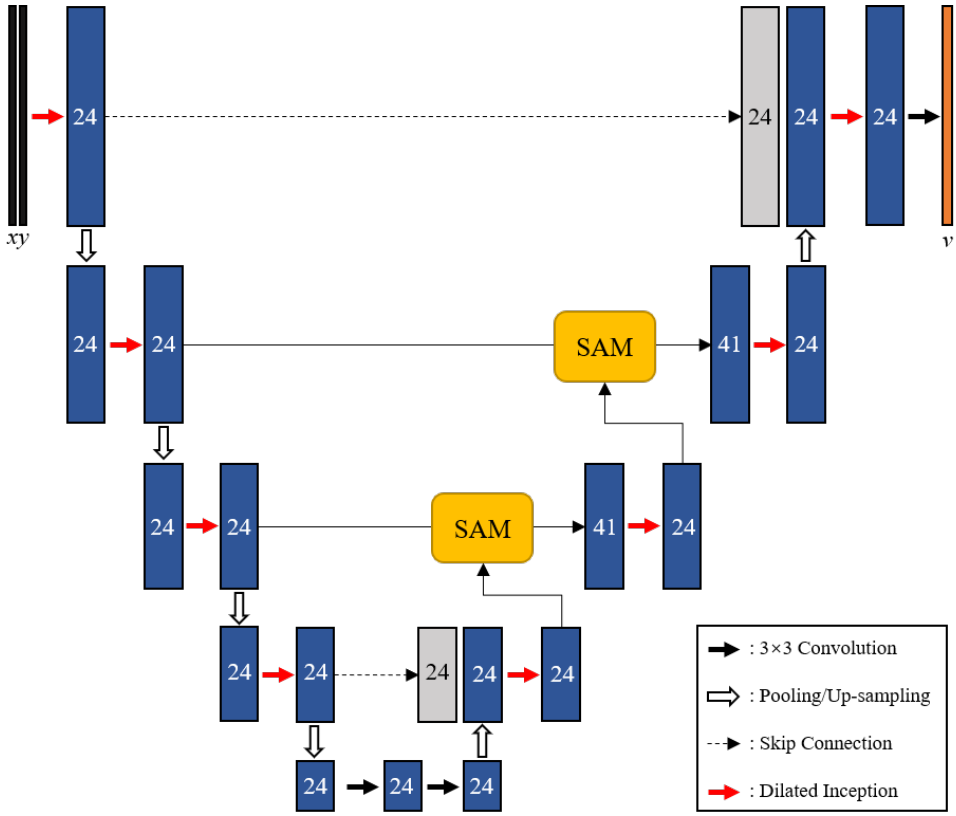

Figure 2.2: Architecture of the scale-adaptive registration network. Numbers inside the blocks indicate number of channels.

Figure 2.2 shows the architecture of the networks. In both of the encoding and decoding paths, most of the standard 3×3 convolution layers in the conventional U-net are replaced with DIMs to increase the receptive field size (details in Sec. 2.3.3).

In addition, SAMs are introduced into the skip connecting paths at the second (1/2) and third (1/4) resolution levels to better extract shallow features using kernels with adaptive receptive fields learned from deep features (details in Sec. 2.3.4).

All the standard 3×3 convolution layers use a stride of 1, zero-padding, and ReLU activation, except the last layer, which uses linear activation. Average pooling and bilinear interpolation with a scaling factor of 2 are used in the down- and up-sampling layers, respectively.

2.3.3 Dilated Inception Module (DIM)

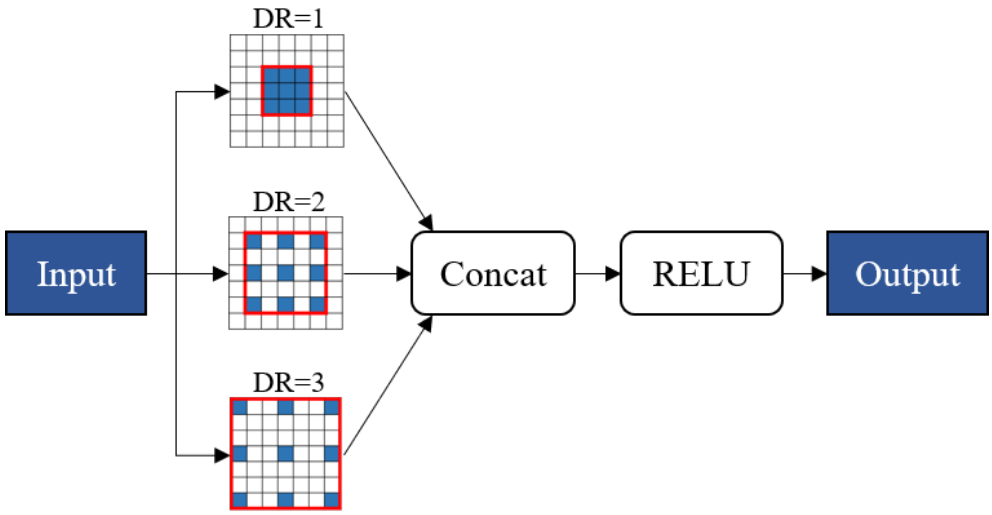

Figure 2.3: Dilated inception module.

Dilated convolution, also known as atrous convolution, enlarges the receptive field by inserting spacing into convolutional kernel according to the dilation rate (DR) [YK15, CPK17]. A 2D dilated convolution operation can be written as:

$$g[a,b] = \sum_{m=1}^{M}\sum_{n=1}^{N} f\left[a + r\cdot m, b + r\cdot n\right]\cdot h[m,n], \tag{2.4}$$

where f is the input feature map, g is the output feature map, h is the convolutional kernel with effective kernel size $M \times N$, and r is the dilation rate. m and n, and a and b are the spatial indices for the convolutional kernel and feature map, respectively. When $r = 1$, the operation reduces to a standard convolution layer. Compared to standard convolution with a large kernel size, dilated convolution is able to achieve large receptive fields without additional network parameters. Compared to using a large (> 1) stride in a convolution layer, dilated convolution keeps the spatial resolution and may preserve spatial information better.

Inception blocks have been used for classification and detection tasks in computer vision and were shown to improve performance with a modest increase in computational

cost [SLJ15]. An inception block combines convolutional operations with multiple kernels of different sizes, and concatenate their output features to form a single output to be fed into the next layer. The output of the inception block contains features with effective receptive fields of different sizes.

We propose to integrate dilated convolution kernels into an inception block to combine the benefit of computational efficiency in dilated convolution and the richness in scale of the learned features from the inception setup. We refer to this design as a dilated inception module (DIM). As shown in figure 2.3, three 3×3 dilated convolutions (*i.e.*, $M = N = 3$) with dilation rates $r = 1, 2, 3$ are apply to the input feature map in parallel to extract features with different receptive fields. The convolution results are then concatenated and activated with a ReLU function to form the output. In this work, we use 8 channels for each of the 3 kernels, resulting in an output feature map with 24 channels.

2.3.4 Scale Adaptation Module (SAM)

In conventional U-net, shallow features in the encoding path are passed to the decoding path through skip connections. Here we proposed to further utilize the contextual information provided by the deep features to guide the propagation of shallow features in the skip connections according to local scales. We propose to use a scale adaptation module (SAM) to integrate the features in a self-adaptive way.

As shown inside the dashed box in figure 2.4, the input feature map from the previous resolution level in the decoding path is first up-sampled by 2 and passed through a standard convolution layer. It is then mapped to the range of $(1, 3)$ with $\tanh(\cdot)+2$ as an activation function to form the corresponding estimate of dilation rate map (DRM). Each pixel value in the DRM is used to assign the dilation rate of a 3×3 dilated convolutional kernel at each corresponding spatial location. This convolution with spatially adaptive dilation rate is applied to the input features from the encoding path to generate a feature map with adaptive scale. Finally, this feature map is concatenated with the up-sampled decoding

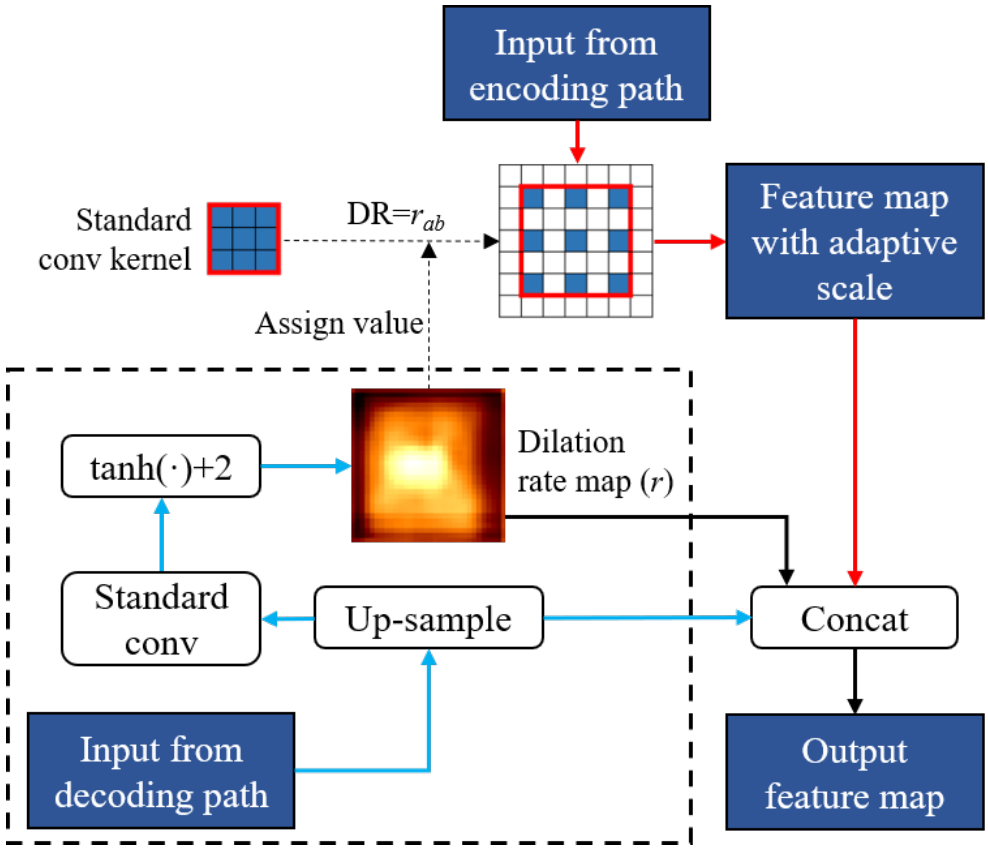

Figure 2.4: Scale adaptation module. "Input from decoding path" is from the lower resolution level, and its path within the module is indicated by the blue arrows. "Input from decoding path" is via the skip connection in the U-net, and its path within the module is indicated by the red arrows.

features and the DRM to form the output of SAM. The up-sampled decoding features, the adaptive dilated convolution, and the DRM have 24, 16, and 1 channels, respectively, yielding an overall output feature map of 41 channels (as illustrated in figure 2.2). The standard convolution layer for DRM generation uses kernel size of 3×3 for the SAM at the 1/2 level and 5×5 at the 1/4 level.

The implementation of the adaptive dilated convolution is slightly different from equation (2.4). Since the dilation rate is limited within the range of $(1, 3)$, we first compute the feature maps from dilated convolutions with $r = 1, 2, 3$, respectively. Note that the parame-

ters for the three dilated convolutional kernels are shared because they arise from the same standard kernel. For fractional dilation rates, linear interpolation of the three feature maps at each spatial location is used as a surrogate:

$$g(r) = t_1 \cdot g(1) + t_2 \cdot g(2) + t_3 \cdot g(3), \tag{2.5}$$

where $t_1 + t_2 + t_3 = 1$, numbers inside the parentheses indicate dilation rates.

The concept of adaptive dilation rate has been investigated by a few studies in semantic segmentation and medical image segmentation [ZTZ17, ZZL20, GCY20]. However, to the best of our knowledge, this is the first study that introduces adaptive dilation rate into a DIR network or into a 3D setup. Our approach of the adaptive dilated convolution operation is faster and requires much less memory than the approach presented by Guo et al. [GCY20], where the bilinear interpolation of the feature map involves a large interpolation array, whose high dimensionality makes it computationally inefficient. The reduction in memory usage enables the 3D implementation of our method.

2.4 Experiments and Results

The network was implemented in both 2D and 3D using TensorFlow. The proposed method was compared against a conventional B-spline-based method in SimpleElastix [MBS16] and three simplified versions of the registration network: U-net alone, U-net with DIM, and U-net with SAM. The objective function was kept the same all methods. In SimpleElastix, a multi-resolution strategy was used, with 20 optimization iterations in each of the four resolution levels. The experiments were performed on a workstation equipped with an NVIDIA GTX 1080 Ti GPU and an Intel i7-6700HQ 3.5 GHz CPU.

2.4.1 Experiment on Cardiac MRIs

2.4.1.1 Data

The method was tested on 2D cardiac MRI sequences obtained from Sunnybrook Cardiac Data [RLC09], which contains 45 4D short-axis cardiac cine MR scans, each containing 20 frames that cover the cardiac cycle. The image resolution was 256×256, with 10 slices, pixel spacing 1.25 mm, and slice thickness 8 mm. Segmentations of left ventricular cavity was provided in the dataset at end-diastole (ED) and end-systole (ES) frames. 45 4D scans were divided into training, validation, and testing sets, containing 30, 5, and 10 scans, respectively. Fixed and moving image pairs were prepared by picking 2D slices from the same 4D scan, at the same slice position but at different time points in the cardiac cycle. Down-sampled in the temporal domain, 27,000 2D image pairs were used for training eventually. A typical image pair is shown in figure 2.5 (a,b).

2.4.1.2 Network Training

Each registration network was trained in mini-batches of 8 image pairs for 80 epochs. The balancing weight λ in equation 2.1 was set to 2 as a result of tuning. ADAM optimizer with a learning rate of 10^{-4} was used.

2.4.1.3 Evaluation

Using the provided segmentations of left ventricular cavity, we computed the following metrics to evaluate the registration performance.

Dice coefficient [Dic45, YV18] between the propagated segmentation mask $M_{x'}$ and the segmentation masks on the fixed image M_y:

$$Dice = \frac{2|M_{x'} \cap M_y|}{|M_{x'}| + |M_y|}. \tag{2.6}$$

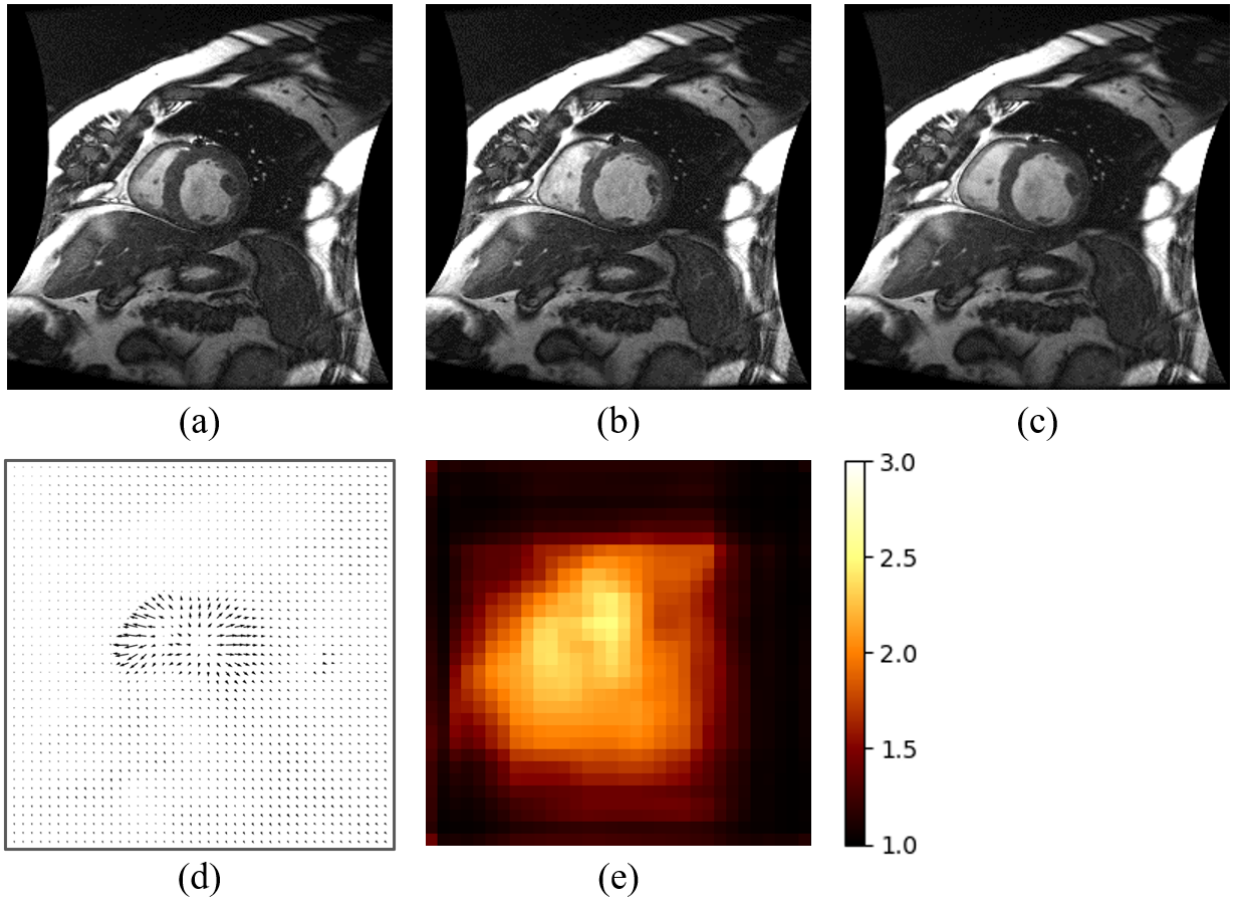

Figure 2.5: Example cardiac MRI registration result in our method. (a) Moving image. (b) Fixed image. (c) Warped image. (d) DVF. (e) Visualization of the DRM at the 1/4 level.

Average surface distance (ASD) [YV18] between the propagated and fixed segmentation contours:

$$ASD = \frac{\sum_{p_{x'} \in C_{x'}} dist(p_{x'}, C_y) + \sum_{p_y \in C_y} dist(p_y, C_{x'})}{|C_{x'}| + |C_y|}, \tag{2.7}$$

where $p_{x'}$ and p_y are points on the propagated contour $C_{x'}$ and the contour on the fixed image C_y, respectively.

2.4.1.4 Results

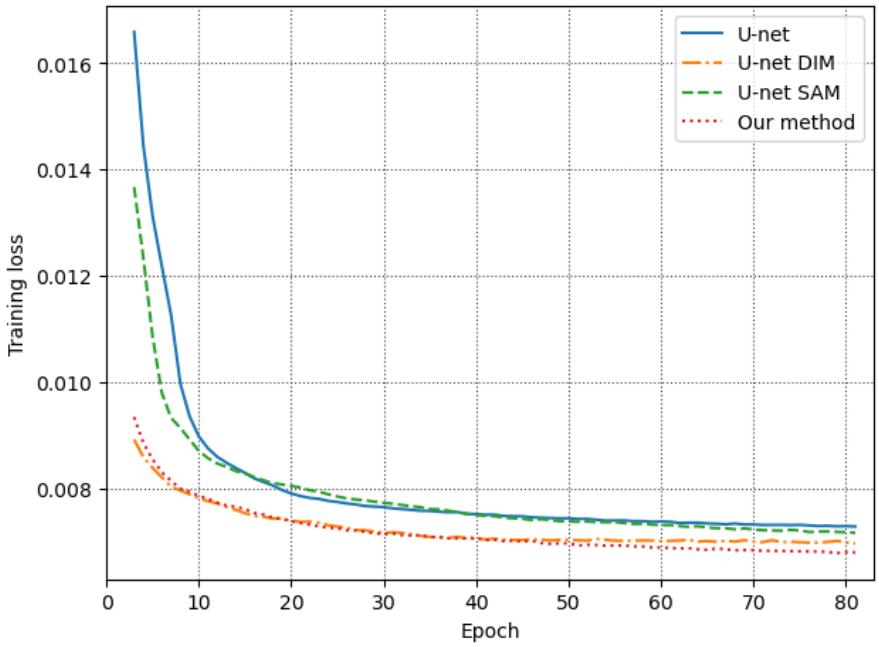

Figure 2.6: Learning curves showing the total loss during the training of the networks. Values for the first two epochs are not displayed.

Figure 2.6 shows the learning curves of the networks. It can be observed that both DIM and SAM effectively reduced the final loss value and DIM provided more significant improvement. The self-adaptation mechanism in SAM made the networks slightly slower to converge. At around the 40th epoch, SAMs began to surpass the counterparts without SAM.

Figure 2.5 shows an example registration result. It can be observed that large dilation rate values in the DRM in (e) correspond to the spatial locations with large deformation in (d).

Table 2.1 shows the quantitative results. Wilcoxon signed-rank tests were performed to examine the statistical significance (at significance level of 0.05, Z_{crit}=1.96). In terms of dice coefficient, our method achieved the best result among all the methods tested, with

statistical significance. In terms of ASD, our method outperformed other networks and was comparable to the best-performing SimpleElastix results without statistical significance. The introduction of SAM increased the network registration time by 2 milliseconds, but the network still offers more than 3 orders of magnitude speed-up over SimpleElastix.

Table 2.1: Results of the cardiac MRI registration experiment. Results are provided as mean ± standard deviation (Z-value from Wilcoxon signed-rank test). Z-values that indicate statistical significance are underlined.

	Dice	ASD (mm)	Time (s)
SimpleElastix	0.92±0.03 (<u>2.13</u>)	**1.65±0.98** (1.89)	5.63
U-net	0.91±0.04 (<u>4.72</u>)	1.73±1.37 (<u>3.11</u>)	**0.002**
U-net DIM	0.93±0.02 (<u>2.00</u>)	1.69±1.01 (1.60)	**0.002**
U-net SAM	0.91±0.03 (<u>3.56</u>)	1.73±1.33 (<u>2.95</u>)	0.004
Our method	**0.93±0.02**	1.68±0.95	0.004

To further appreciate the performance of the SAM, we calculated histograms of the dilation rates in the SAM at the 1/4 resolution level. A comparison was made between DRMs for image pairs of (a) adjacent frames near ED, where deformation is small, and (b) ED and ES frames, where deformation is large, as shown in figure 2.7. It can be observed that the distribution of the estimated dilation rates is in agreement with the qualitative motion levels: with the larger deformations between ED and ES taking a larger proportion of high dilation rate. The proportion of dilation rate values larger than 2.25 $P(r > 2.25)$ was 0.023 and 0.088 for the small and large deformations, respectively.

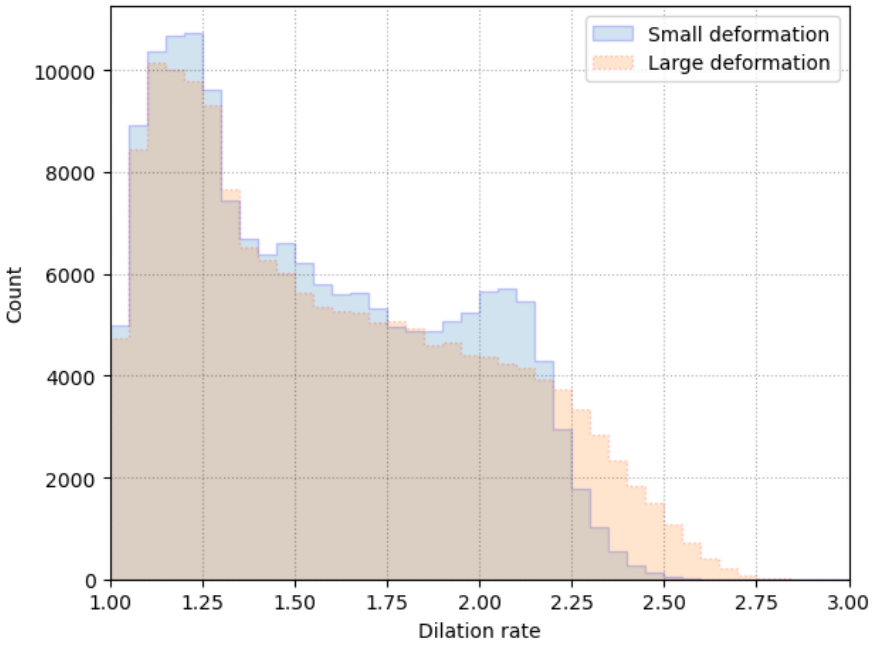

Figure 2.7: Histograms showing the distributions of dilation rates in the SAM for testing image pairs with small and large deformations. The number of pixels are the same, and the total counts are equal for the two input settings.

2.4.2 Experiment on Synthetic Data

2.4.2.1 Data Generation

Given the absence of ground-truth DVFs in clinical images, we synthesized DVFs and images as digital phantoms for a quantitative evaluation on dense DVFs. Data generation was based on the same dataset as described in Sec. 2.4.1.1. As shown in figure 2.8, first we used SimpleElastix (with the same objective function and setup) to generate DVFs between ES and ED frames of the images. These DVFs were multiplied by a scaling factor α uniformly distributed within the range $[0.5, 1.2]$ for data augmentation. The scaled DVFs were taken as ground truths and were used to generate the warped images in the forward direction with corresponding moving images. Training, validation, and testing sets contained 4800, 800,

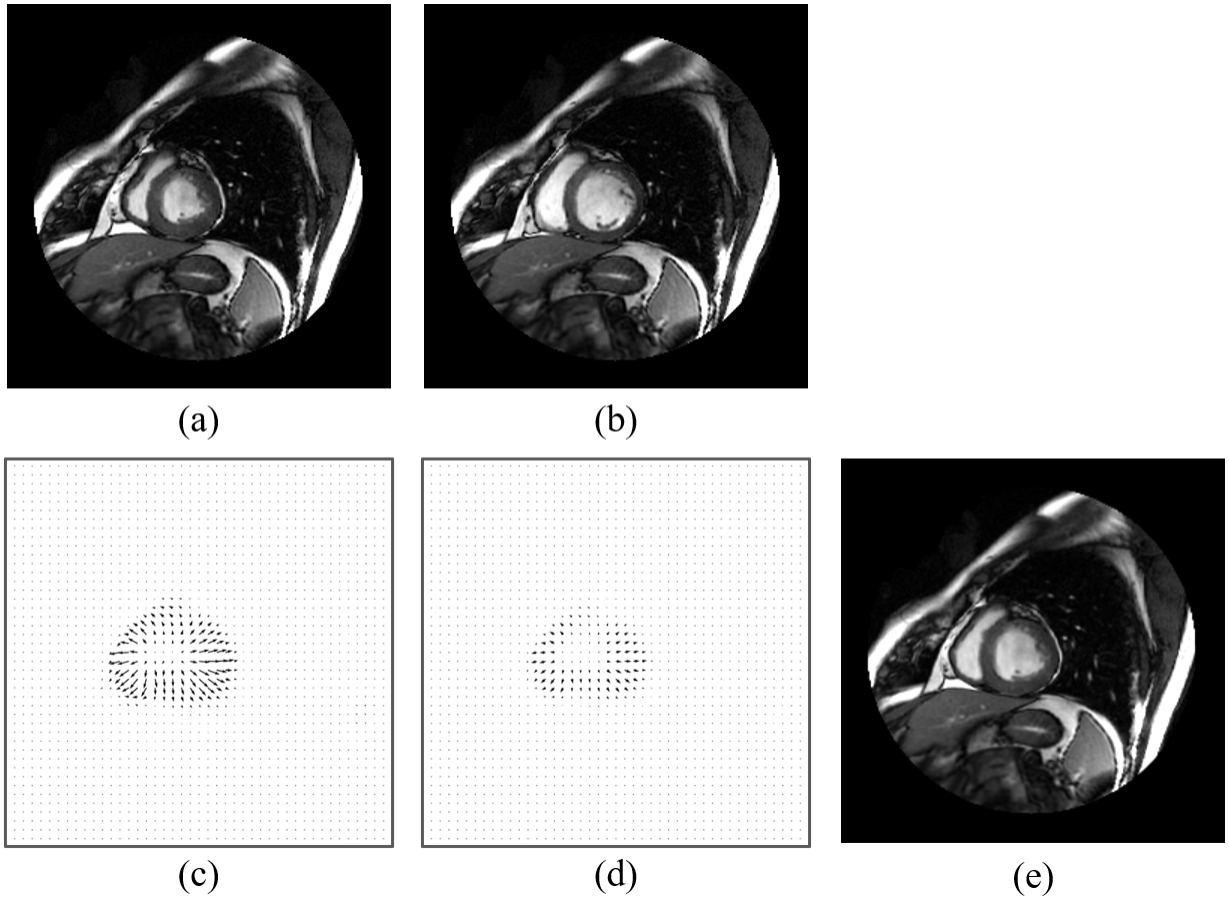

Figure 2.8: Example synthetic data. (a) Moving image. (b) Fixed image in SimpleElastix. (c) SimpleElastix DVF. (d) Scaled DVF with $\alpha = 0.6$. (e) Forward generated fixed image.

and 1600 2D synthetic samples, respectively.

2.4.2.2 Network Training

Each registration network was trained in mini-batches of 8 image pairs for 500 epochs. The balancing weight λ in equation 2.1 was set to 2. ADAM optimizer with a learning rate of 10^{-4} was used.

2.4.2.3 Evaluation

Root mean squared error (RMSE) to the ground-truth DVF was calculated as a dense version of target registration error ($\mathrm{TRE_d}$) to indicate registration performance:

$$TRE_d = \sqrt{\frac{1}{d_a d_b} \sum_{a=1}^{d_a} \sum_{b=1}^{d_b} \|v^*_{ab} - v_{ab}\|^2}, \tag{2.8}$$

where v^* and v are the ground-truth and result DVFs.

2.4.2.4 Result

Paired t-tests were performed to examine the statistical significance. As shown in table 2.2, our method achieved the lowest TRE among all the networks, with statistical significance.

Table 2.2: Results of the synthetic image registration experiment. Results are provided as mean ± standard deviation and p-value from paired t-test.

	$\mathrm{TRE_d}$ (mm)	p-value
U-net	0.024±0.020	2.75×10^{-8}
U-net DIM	0.019±0.011	9.53×10^{-3}
U-net SAM	0.022±0.020	2.12×10^{-7}
Our method	**0.017±0.009**	-

To further appreciate the performance of the SAM, we examined the DRM at the 1/4 resolution level and calculated $P(r > 2.0)$ and $P(r < 1.5)$ for varying magnitudes of deformation introduced by varying α values. The results are shown in figure 2.9. With larger deformation magnitude quantified by α, the proportion of small dilation rate decreased and the proportion of large dilation rate increased.

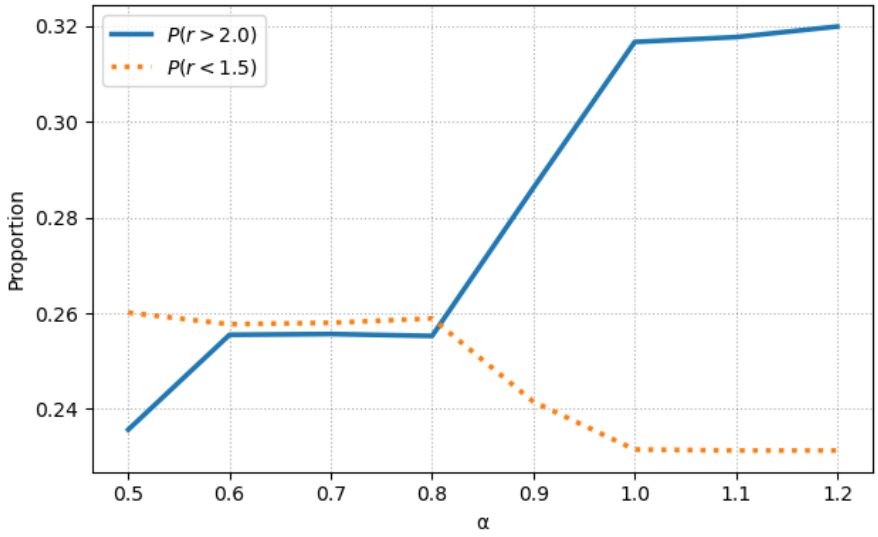

Figure 2.9: Statistics of the dilation rates for varying magnitudes of deformation introduced by varying α values for data synthesis.

2.4.3 Experiment on lung 4DCT

2.4.3.1 Data

We also evaluated the 3D version of the network using two publicly available thoracic 4DCT datasets: 4D-Lung collection from the Cancer Imaging Archive (TCIA) [HWS17] and DIR-Lab [CCG09, CCM09]. The 4D-Lung collection consists of scans acquired during chemoradiotherapy of 20 locally-advanced, non-small cell lung cancer patients. One scan from each patient was used. The images were acquired on a 16-slice helical CT scanner as respiration-correlated CTs with 10 breathing phases. The reconstructed slice thickness was 3 mm and in-plane spacing was 0.98 to 1.17 mm. The DIR-Lab dataset consists of 10 scans acquired as part of the radiotherapy planning process for the treatment of thoracic malignancies. The images also have 10 breathing frames. The slice thickness was 2.5 mm and in-plane spacing was 0.97 to 1.16 mm. Each scan contains 300 manually identified anatomical landmarks annotated at end-exhale (EE) and end-inhale (EI) phases.

The 4D-Lung collection was divided into training and validation sets, containing 15 and 5 scans, respectively. 1350 intra-subject 3D image pairs were used for training eventually. The DIR-Lab dataset was used as an independent testing set to realistically reflect the predicted performance on images acquired from different institutions and machines. All images were resampled with slice thickness 2.5 mm and typical in-plane pixel spacing 1.16 mm and then cropped with a $256 \times 256 \times 96$ window that covered the lungs. Image intensities were clamped between -1000 and 500 HU and scaled between 0 and 1.

2.4.3.2 Network Training

Each registration network was trained for 40 epochs with a batch size of 1. The balancing weight λ in equation 2.1 was set to 0.5. ADAM optimizer with a learning rate of 10^{-4} was used.

2.4.3.3 Evaluation

The 3D Euclidean distance between transformed and fixed anatomical landmarks was calculated as target registration error (TRE_l) to indicate registration performance.

2.4.3.4 Results

Figure 2.10 shows an example registration result. Note that the 3D DVF is visualized with its 2D projection on the coronal plane. The large (> 2.25) dilation rate values in the DRM concentrated around the diaphragm, which was the major driving force for the respiratory motion and had the largest motion magnitude. While the subdiaphragmatic regions also had relatively large deformations, we observed that they did not necessarily correspond to large dilation rate values.

As shown in table 2.3, our method achieved the best TRE among all the methods tested, with statistical significance (under paired t-test with $p < 10^{-6}$).

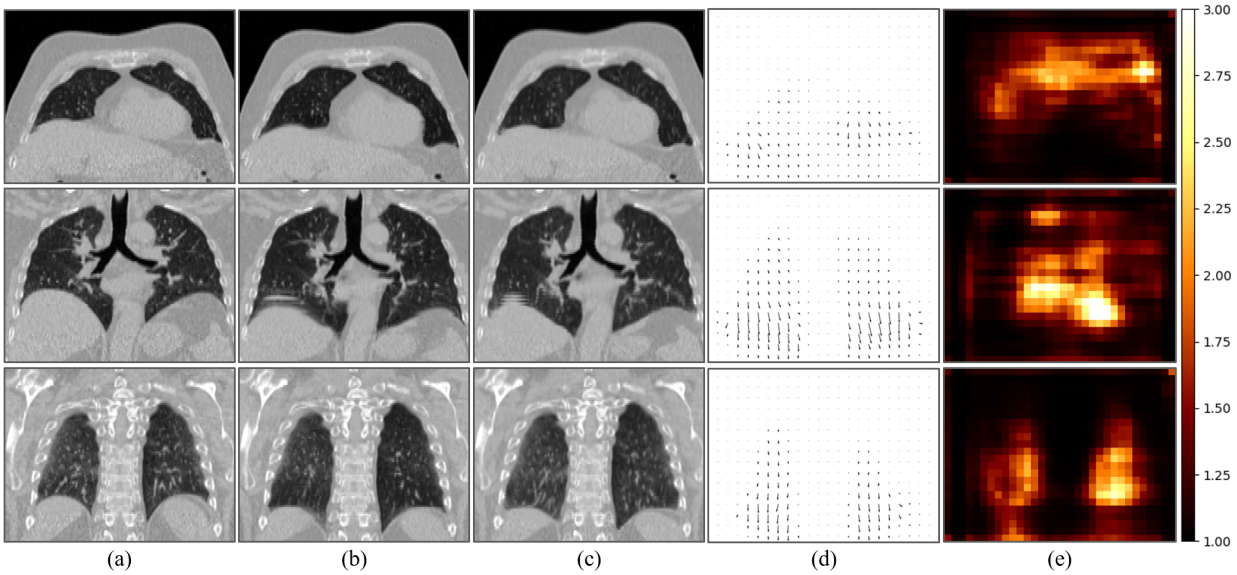

Figure 2.10: Example thoracic CT registration result in our method. Three coronal slices of a 3D volume is shown. (a) Moving image. (b) Fixed image. (c) Warped image. (d) In-plane components of DVF. (e) Visualization of the DRM at the 1/4 level.

2.5 Discussion

In the SAMs, we used linear interpolation of pre-computed feature maps with integer dilation rates as computationally efficient surrogates to convolutional kernels with fractional dilation rates. A similar SAM idea has been utilized in segmentation by Guo et al. [GCY20], where bilinear interpolation of the feature map was used and involved explicit manipulation of a large interpolation array. Its high dimensionality requires high memory capacity and is computationally inefficient. A 2D SAM development with 128×128 DRM easily drives the explicit interpolation out of memory in our setting with 11 GB GPU memory. Our approach of the adaptive dilated convolution operation is more memory-economical and faster, enabling a 3D implementation on the same lower budget platform. To the best of our knowledge, this is the first study that introduces adaptive dilation rate into a DIR network or into a 3D setup.

Table 2.3: Results of the lung CT registration experiment. TREs are provided as mean ± standard deviation in millimeter.

	Initial	SimpleElastix	U-net	U-net DIM	U-net SAM	Our method
Case 1	3.89±2.78	1.27±1.13	**1.26±0.71**	1.40±0.78	1.30±0.71	1.27±0.80
Case 2	4.34±3.90	1.44±1.28	1.69±1.28	1.34±0.79	1.50±1.01	**1.26±0.84**
Case 3	6.94±4.05	1.77±1.37	2.25±1.56	1.73±1.11	1.93±1.27	**1.55±1.14**
Case 4	9.83±4.85	2.26±1.88	3.32±2.26	2.33±1.64	2.80±1.88	**2.10±1.66**
Case 5	7.48±5.50	2.67±2.70	2.79±2.30	2.50±1.84	2.59±1.97	**2.30±1.98**
Case 6	10.89±6.96	3.82±2.99	3.36±2.07	2.84±2.07	3.18±2.18	**2.39±1.53**
Case 7	11.03±7.42	3.69±3.15	5.02±3.51	3.40±2.64	4.22±3.01	**3.08±2.48**
Case 8	14.99±9.00	7.03±8.25	8.81±6.99	**5.52±5.83**	6.98±6.15	6.01±6.65
Case 9	7.92±3.97	2.43±1.61	3.45±2.10	2.76±1.55	3.11±1.76	**2.42±1.55**
Case 10	7.30±6.34	3.21±2.89	3.48±3.12	3.05±2.76	3.24±2.92	**2.80±2.53**
Total	8.46±6.58	2.96±3.73	3.54±3.68	2.69±2.78	3.09±3.12	**2.52±2.96**
Time(s)	-	46.50	**0.12**	**0.12**	0.42	0.42

In the 2D experiments, DRMs at the 1/2 level almost always converged to a uniform map of 1 (i.e., standard convolution), regardless of the kernel initialization methods, presence or removal of the DIMs, etc. One possible explanation was that features at the 1/2 level were close to the output layer and focused more on the demand of using local scale to provide sufficient resolution drive for DVF refinement. Although such kernels converged to standard convolutions did not hinder the overall network performance, we plan to further investigate this problem and determine the optimal SAM placement.

In the 3D experiments, we observed that the subdiaphragmatic regions with relatively large deformations did not necessarily correspond to large dilation rate values, as shown in figure 2.10. One possible reason is that local textures were missing in those tissues, especially

after clamping the image intensity, and DVF estimation in those regions was mainly driven by features on an even more global scale.

The DIR-Lab dataset described in Sec. 2.4.3.1 has also been used to evaluate the multi-stage deep learning DIR presented by de Vos et al. [VBV19], reporting a landmark-based TRE of (2.64 ± 4.32) mm. This error was larger than our method and smaller than the three simplified versions of our network. However, when comparing the TRE results, it should be noted that de Vos et al. performed a leave-one-out cross-validation on the 10 images in DIR-Lab, while in our experiment, the networks were trained on 15 images from an independent TCIA source and were tested on DIR-Lab. Our result was based on a larger training set, but was challenged by the image variations across datasets and more frequent motion artifacts in the 4D-lung collection.

Our method was evaluated with three intra-subject experiments because the adaptation of scale can be better demonstrated and appreciated with the time-resolved motions. We expect the advantage of adaptive scale to translate to inter-subject registration, which also need strong accommodation for deformation across various scales.

2.6 Conclusion

In this study, we have presented a deep neural network for DIR. The major contribution of this work is the introduction and integration of DIMs and SAMs, which address the heterogeneous scale problem with self adaptation and high efficiency in both GPU memory utilization and computation time. The self-adaptive dilation rate is in sharp contrast to CNN architectures using a fixed kernel size in each convolution layer, which has to be prescribed a priori. The DIMs explicitly enlarge the effective receptive field without additional network parameters. The SAMs process the shallow features from the encoding path through skip connections guided by deep decoding features and allow network parameters to be updated based on the spatial region that covers the local deformation scale. The effectiveness of

the modules was shown in the experiments. Our method achieved better or comparable results compared to classic hierarchical B-spline methods in SimpleElastix, where the scale heterogeneity was addressed with a multi-resolution strategy.

CHAPTER 3

Incorporating Feasibility Prior into Deformable Image Registration

3.1 Introduction

The class of allowable deformations is determined by the choice of parametrization models such as piecewise affine transform, B-splines, thin-plate splines, etc. Such explicit parametrization usually still renders highly under-determined registration problems, resulting in instability of solutions and local optima issues. Regularizations are therefore introduced to alleviate these issues and incorporate prior knowledge into the problem formulation. The regularization terms usually encourage physically or physiologically feasible deformations, and quantitatively describe the deviation from smoothness, diffeomorphism, etc.[SDP13, Hol07].

One major challenge in regularization design lies in the mathematical quantification of physical and physiological properties. Despite the long efforts in designing models and regularizers for invertibility, diffeomorphism, volume preservation and interface discontinuity preservation [JC02, Ash07, RFR06, REF09, PHF14], an ideal solution remains elusive. At the center of this challenge is spatial heterogeneity. Tissue properties including elasticity and anisotropic discontinuity or sliding, vary across the domain of interest. This indicates that spatially adaptive local orders and balancing weights are necessary for parametric models and regularizations, respectively, which is beyond the descriptive power of existing registration methods with a few global hyper-parameters.

The recent development of deep learning methods provides an alternative to the conventional registration framework. In the context of registration, a deep network can be trained to infer DVFs directly from input image pairs. Training of the network can be either supervised [YKS17, CYZ18, SVB17, KMD17] or unsupervised [VBV17, VBV19, BZS19, FCW19, LF18, Zha18, DBG18, WAH20, HGG18, SR20, YXR18, MAS18].

In the supervised learning methods, typically a deep neural network is trained with ground-truth DVFs corresponding to the image pairs in the training set. Unfortunately, true DVFs are usually inaccessible, and even manually generated flows are prone to large errors and uncertainty. The usual approach for creating legitimate training samples is by generating DVFs from image pairs using other registration approaches [YKS17, CYZ18], or by simulating DVFs first and computing the warped images afterward [SVB17, KMD17]. The corresponding DVF feasibility prior depends on the quality of the "ground truth" and couples with the registration [YKS17, CYZ18] or simulation approach [SVB17, KMD17], and limits the overall performance. Despite the benefit of efficiency, it is difficult for these methods to outperform state-of-the-art optimization-based approaches.

In the unsupervised learning methods, a spatial transformation module as part of the network architecture enables the computation of warped images and image similarity in the training process. With explicit loss functions, they are likely to have similar behaviors as the conventional methods, but the challenge of model and regularization design remains unsolved. In principle, unsupervised DIR is compatible with any differentiable regularizer from the classic image-based methods. Existing investigations have been mostly generic without much tailoring to the respiratory motion so far. DVF smoothness is typically encouraged using the first or second order spatial gradients [VBV19, BZS19, Zha18, FCW19, LF18, JYG20, FB20, FLW20a]. Zhang introduced inverse-consistent and anti-folding constraints by interchanging the fixed and moving images and inverting the DVFs in the training setup [Zha18]. Dalca et al. imposed diffeomorphic constraints into registration network using a stationary velocity field representation defined via ordinary differential equation (ODE)

integration in the scaling and squaring layers [DBG18]. Wei et al. further introduced a tissue-aware Jacobian determinant regularizer to the diffeomorphic DVF after the scaling and squaring operations to avoid folding and non-smoothness [WAH20]. Hu et al. introduced adversarial deformation constraints with a discriminator network, which distinguishes the registration-predicted displacement fields from the motion data offered by biomechanical models [HGG18].

In this study, we propose two different approaches to impose feasibility conditions on DIR. Both approaches are based on building implicit feasibility prior on DVF. In the first approach, a supervised feasibility model is trained from a set of physiologically reasonable DVFs using a convolutional auto-encoder (CAE). The CAE is used as a flexible regularizer when training the DIR network. In the second approach, an unsupervised feasibility model is trained from moving and fixed image pairs directly using a statistical generator network without any pre-generated DVF sample. It is then used as a novel DVF parametrization model when performing DIR.

3.2 Supervised Feasibility Prior Based on Convolutional Auto-encoder

3.2.1 Method

3.2.1.1 Overview

Our method consists of two major modules developed sequentially. First, a feasibility descriptor in the form of a CAE is trained in a supervised setting using a set of physically reasonable DVFs derived from high-quality images to characterize feasible respiratory motions. Then, an unsupervised DVF estimation network is trained with potentially low-quality images, which includes the trained CAE as a Plug-and-Play (PnP) regularizer to simultaneously enforce image matching and penalize DVFs that deviate from the learned implicit

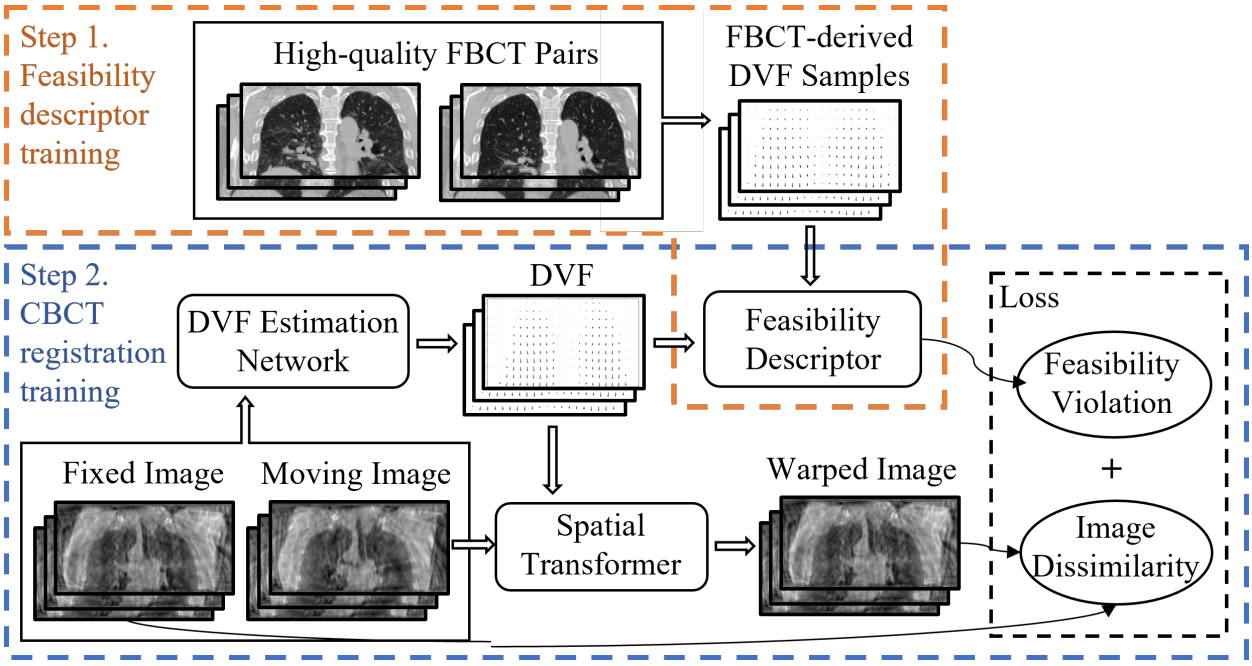

Figure 3.1: Overview of the proposed CAE-based method. Our method consists of two steps. In the first step, DVFs derived from high-quality images are used to train a feasibility descriptor to capture the underlying feasibility manifold. In the second step, the feasibility descriptor is incorporated into an unsupervised DIR network to regularize the estimated DVFs.

feasibility condition.

As shown in Fig. 3.1, the proposed method consists of a DVF estimation network, a spatial transformer [JSZ15], and feasibility descriptor. The major contribution of this work is the injection of the feasibility descriptor, which is trained independently beforehand and then plugged into the DVF estimation network. Loss functions defined as the combination of the dissimilarity between the fixed and warped images and the DVF feasibility violation provided by the descriptor, are used to drive the network optimization, with a back-propagation scheme.

3.2.1.2 DVF Feasibility Descriptor

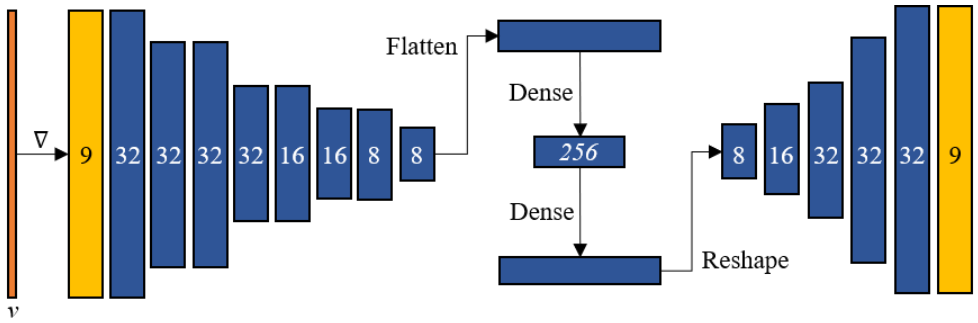

Figure 3.2: Architecture of the feasibility descriptor.

To introduce spatially variant regularization on deformation, we use a CAE to model the DVF feasibility conditions. As shown in Fig. 3.2, the Jacobian matrix of an input DVF is computed and represented by a nine-channel feature map. In the encoder, four alternating layers of convolution and average-pooling are applied. Then, a fully connected layer with 256 units is used, which generates the latent representation of the DVF. The decoder path consists of a fully connected layer and three transposed convolution layers with stride of 2. All the (de)convolution layers use zero-padding, kernel size of 3, and ReLU activation, except the last layer, which uses linear activation.

The loss function used for the CAE training is the squared Frobenius norm of point-wise difference between the DVF Jacobian and the CAE output:

$$L_V = \| \nabla(v) - \mathrm{CAE}(\nabla(v)) \|^2_{\mathrm{Frob}}, \tag{3.1}$$

where v is a DVF sample, and ∇ is the Jacobian operator. Once the CAE is properly trained, the deviation of a candidate DVF from the manifold, as described by the auto-encoding discrepancy of its Jacobian matrix in squared Frobenius norm, provides a measure of the physical or physiological feasibility of that DVF. With the CAE trained under supervision with site-specific data, we expect it to consolidate spatial variant smoothness and directional-specific characteristics, and transfer such knowledge into the subsequent DVF estimation

network.

3.2.1.3 Unsupervised Learning for DVF Estimation

The DVF estimation network takes concatenated pairs of moving and fixed images as input, and outputs a DVF. As shown in Fig. 3.3, the network uses a general U-net structure [RFB15] to take advantage of the hierarchical structure and skip connections for effective learning of features at all scales. All the $3 \times 3 \times 3$ convolution layers use a stride of 1, zero-padding, and ReLU activation, except the last layer, which uses tanh activation. Average pooling and up-sampling with a scaling factor of 2 are used in the encoding and decoding paths, respectively.

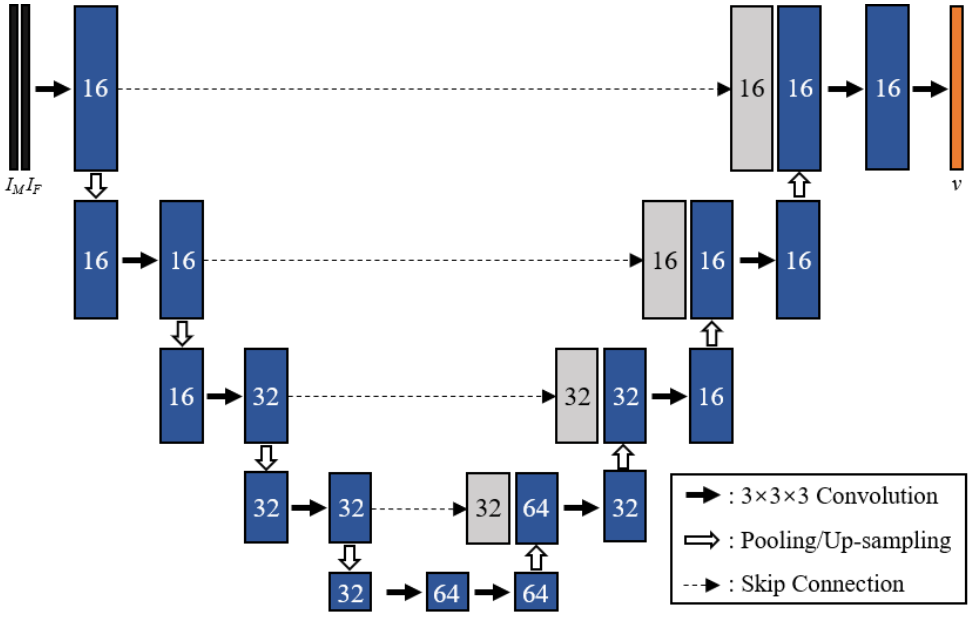

Figure 3.3: Architecture of the DVF estimation network.

The loss function for the DVF estimation network training consists of an image intensity matching cost L_D and an implicit feasibility violation penalty on DVF Jacobian L_V. In this work, we used normalized NCC as the similarity metric. The loss function can be written

as:

$$L = L_D + \mu L_V = -\text{NCC}(I_F, I_M \circ v) + \mu \| \nabla(v) - \text{CAE}(\nabla(v)) \|^2_{\text{Frob}}, \tag{3.2}$$

where $v(I_F, I_M)$ is the DVF ouput from the estimation network, $I_W = I_M \circ v$ is the image warped with DVF v, and μ is a balancing weight.

3.2.2 Experiments on Lung CT and CBCT

3.2.2.1 Training of the Feasibility Descriptor

To obtain the DVF samples that represent realistic respiratory motion, we performed conventional B-spline registration on a set of 10 CT images from the DIR-Lab dataset [CCG09, CCM09]. The slice thickness was 2.5 mm and in-plane spacing was 0.97 to 1.16 mm. All images were resampled with slice thickness 2.34 mm and in-plane pixel spacing 1.16 mm, and then cropped to a $256 \times 256 \times 64$ window that covered the lungs. Image intensities were clamped between -1000 and 500 HU and scaled between 0 and 1.

Classic B-spline registration in SimpleElastix was used to generate DVFs between breathing phases, with bending energy penalty (BP) regularized NCC objective. In order to accommodate spatially variant regularization, we used various values of regularization weights λ ranging from 0.01 to 2, so that the learned manifold could address different local trade-offs. For each of the 10 scans, 15 moving and fixed image pairs were selected. Then, they were augmented by 5 registrations performed with different λ for BP regularization. As a result, 750 DVFs were generated as the training set. The CAE was trained for 200 epochs with batch size 1. ADAM optimizer with learning rate 10^{-4} was used.

3.2.2.2 Training of the DVF Estimation Network

The 4D-CBCT data was from the 4D-Lung collection in the Cancer Imaging Archive (TCIA) [HWS17]. They were acquired during chemoradiotherapy of 20 locally advanced, non-small

cell lung cancer patients. Each scan has 10 breathing phases. The reconstructed slice thickness was 3 mm and in-plane spacing was 0.98 to 1.17 mm. The images were pre-processed to the same size and pixel spacing described in Sec. 3.2.2.1. The training, validation, and testing sets contain 10, 5, and 5 patients, respectively. In the training set, 25 scans from the 10 patients were used for data augmentation purpose.

For each scan, 15 moving and fixed image pairs were selected. The network was trained for 150 epochs with batch size 1. The balancing weight μ in Eq. (3.2) was set to 10^{-6}. ADAM optimizer with learning rate 10^{-4} was used.

In addition, to test the method's sensitivity to the regularization weight μ, the network was also trained in four other settings with $\mu = 10^{-7}, 10^{-6.5}, 10^{-5.5}, 10^{-5}$.

3.2.2.3 Benchmark Methods

The proposed method was compared against a classic B-spline method, a diffeomorphic Demons method [VPP09], and DVF estimation networks trained without regularization, with bending energy penalty (BP) regularization [VBV19], and with a cooperative CAE [BEK19].

Classic BP: The classic B-spline registration was based on the SimpleElastix toolbox [MBS16]. The cost function is the weighted sum of NCC and BP, with the regularization weight tuned for each case to optimize the performance.

Demons: The diffeomorphic Demons method was based on the Insight Toolkit (ITK) [VPP09]. 250 iterations were used. Gaussian smoothing with standard deviation of 1.0 was applied on the DVFs.

U-net and U-net BP: The proposed DVF estimation network was trained with and without BP. The weight for BP was set to 1. Other hyperparameter settings were the same as Sec. 3.2.2.2.

Coop CAE: The closest learning-based feasibility prior to ours is the cooperative autoencoder in [BEK19], where a CAE is also used as a regularizer, similar to this work. It differs

from our proposal in that the CAE and the DIR network are trained jointly so the CAE mainly acts as a dimensionality regularizer, and it acts on the same image inputs as the DIR without any training on DVF set. We trained the estimation network with the initialization phase proposed in the paper and set the regularization weight to 0.5. Other hyperparameter settings were the same as Sec. 3.2.2.2.

3.2.2.4 Evaluation

Landmark evaluations on real CBCTs: On the five clinical 4D-CBCT from TCIA, we manually annotated ten anatomical landmark pairs at the EE and EI phases.

The Euclidean distance between the transformed and the fixed landmarks was calculated as TRE to measure registration performance. Paired t-tests were used to examine statistical significance.

Landmark evaluations on simulated CBCTs: To obtain landmark annotation on a larger scale, we simulated 4D-CBCTs from CT scans in the SPARE dataset [SGL19]. Each phase in the CBCT was simulated independently, using the corresponding 3D CT data. The geometric setup and the number of projections were set to match the protocol in the TCIA acquisition. FDK reconstruction algorithm was then applied using the TIGRE toolbox [FDK84, BDH16]. As shown in Fig. 3.4, the simulated CBCT has similar image quality to the clinical CBCT.

We applied an automatic landmark pair detection algorithm [FWT19] to the original CTs to take advantage of its higher image quality and structural details. The locations of the landmarks were then mapped to the simulated CBCTs. Eventually, nine scans from the dataset were used, each with 100 landmark pairs in the EI and EE phases.

Enhancement experiment: An accurate DIR should be able to identify the motion trajectory of each pixel, and integration along such trajectory can enhance image quality. A simple motion-compensated image enhancement test was performed by collapsing all phases

of the 4D-CBCT according to the estimated DVFs to an arbitrary reference phase and taking the average.

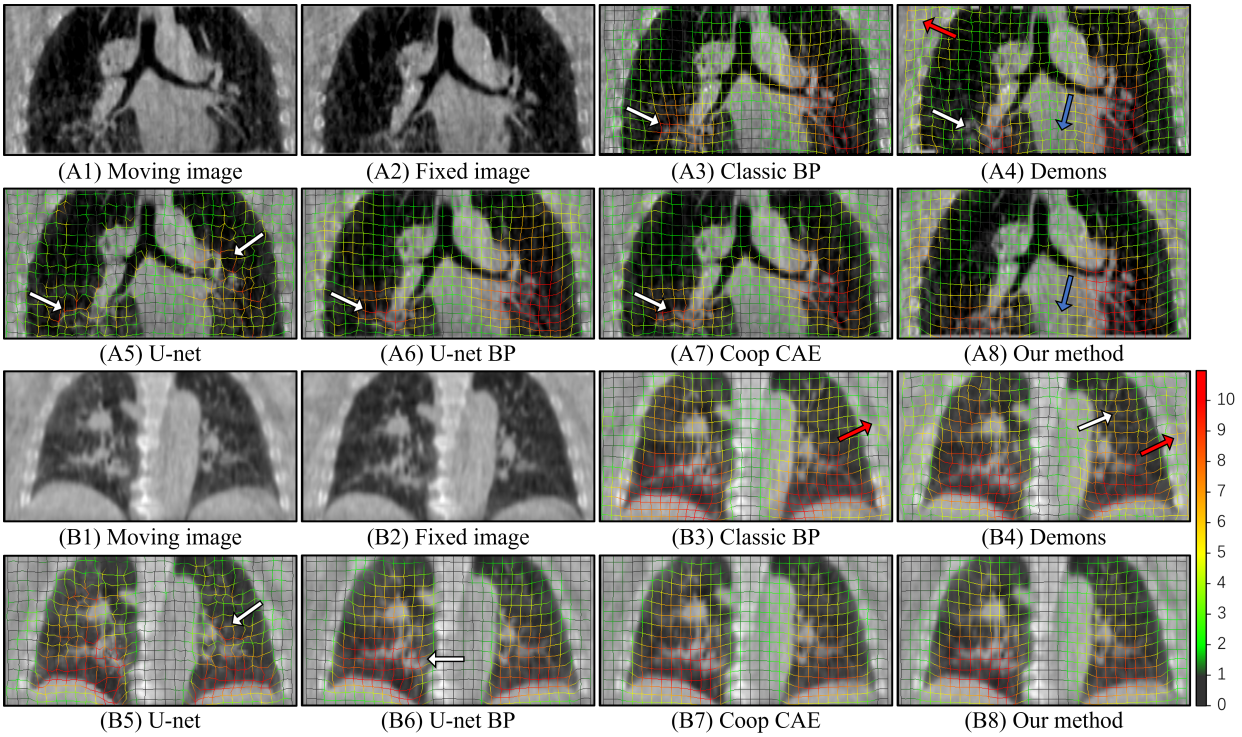

Figure 3.4: Example registration results. The first two rows (A) are from a real CBCT, and the last two rows (B) are from a simulated CBCT. 3D DVFs are visualized with their 2D projection onto the coronal plane with the color indicates the deformation magnitude in 3D (unit: mm). White arrows indicate local non-smoothness. Red arrows indicate dubious motions outside the rib cage.

In the simulated CBCT study, the enhancement is quantified with root-mean-square error (RMSE) and structural similarity index measure (SSIM), in addition to qualitative visualization as with clinical CBCT data.

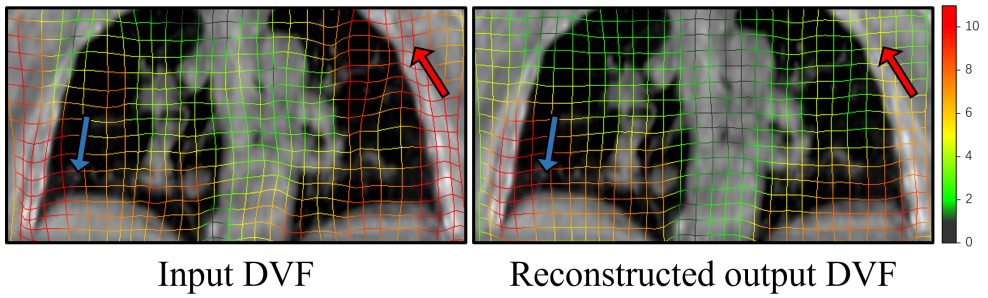

Figure 3.5: Example input DVF to the feasibility descriptor and its reconstructed output. The background is the corresponding fixed image. Color indicates motion magnitude (unit: mm). The CAE reconstruction effectively removes dubious large motion outside of rib cage (red arrows) and preserves physiological large motion driven by diaphragm dynamics (blue arrows).

3.2.2.5 Results

To illustrate the behavior of the CAE-based regularizer, an example input DVF to the trained feasibility descriptor and its corresponding DVF reconstructed from the CAE-output Jacobian are shown in Fig. 3.5. It can be observed that the reconstructed output preserved reasonable motion boundaries but was smoother. The false deformation vectors outside the ribcage were significantly reduced in magnitude, while the large motions close to the diaphragm were preserved. Note that the input DVF is an intermediate result from an incomplete training process, and that the reconstructed DVF is calculated from the output Jacobian through integration, which is not explicitly required as a step in our method, but is illustrated here to demonstrate the behavior and impact of the feasibility descriptor.

Fig. 3.4 shows some example registration results. The desirable local smoothness was most appropriately reflected in the solutions from Coop CAE and our method, as seen in (A8), (B7), and (B8). The small motions outside the rib cage were better estimated by U-net BP, Coop CAE, and our method, whereas other methods were heavily affected by artifacts in that region, as seen in (A4), (B3), and (B4).

Table. 3.1 shows the quantitative TRE results. Our method achieved the best TREs on both real and simulated data. Paired t-tests indicated that the TRE reductions in our method were statistically significant ($p < 0.01$) compared to all the other methods tested. The average registration time was 52 s for classic BP, 82 s for Demons, and 0.04 s for all four networks.

Table 3.1: Target registration errors based on the anatomical landmarks. Results are provided as mean ± standard deviation in millimeter (p-value from paired t-tests).

	Real CBCT	Simulated CBCT
Before	6.12±4.12	7.53±4.15
Classic BP	1.74±1.66 (0.008)	2.55±2.45 (0.001)
Demons	1.98±1.75 (10^{-4})	2.55±2.77 (10^{-4})
U-net	2.45±1.87 (10^{-5})	2.98±3.02 (10^{-8})
U-net BP	1.93±1.62 (0.001)	2.51±2.41 (10^{-6})
Coop CAE	1.68±1.23 (0.008)	2.20±2.01 (0.007)
Our method	**1.63±0.98**	**2.13±1.84**

Fig. 3.6 shows that the model performance is robust with respect to the choice of regularization weight μ, with very minor TRE variations in response to an order of magnitude change in μ.

Fig. 3.7 and Fig. 3.8 show example image enhancement results on real and simulated CBCTs, respectively. Classic BP and Demons still present prevalent streak artifacts after fusion, as these algorithms are driven to register and enhance streaks in the same way as anatomical contents. U-net BP tended to predict smooth motion in the homogeneous region, even when artifacts exist. Since the streak artifacts were associated with the projection angles, which usually differed in each phase, they became less pronounced after averaging multiple phases. Coop CAE and our method imposed stronger prior for respiratory motion and generated better results with less noise and artifacts. Because of the higher accuracy,

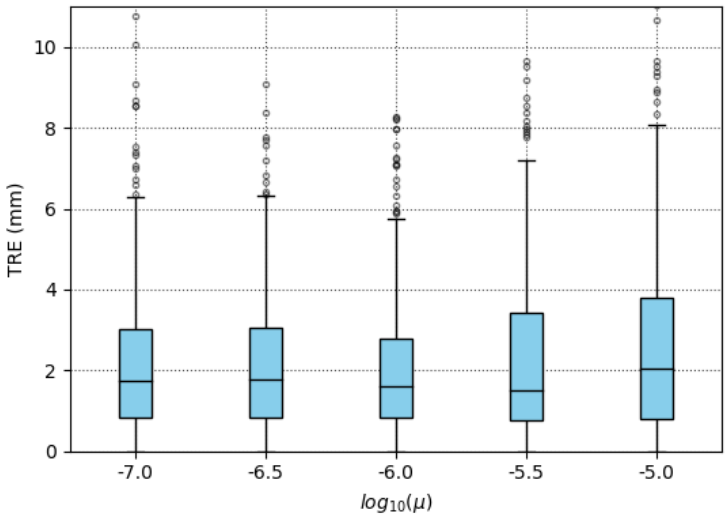

Figure 3.6: Simulated CBCT registration result from networks trained with different balancing weights μ for the feasibility violation loss. The horizontal axis is shown in log scale.

our method achieved higher visual resolution and managed to reconstruct sharper detail structures as indicated by the red ovals in the figures.

Fig. 3.9 shows the image intensity profiles before and after enhancement. Since classic BP achieved a decent result on this particular image shown in Fig. 3.8, its profile is displayed for comparison. Our method had the sharpest transition and closest attenuation match.

Table. 3.2 shows the quantitative results. The RMSE and SSIM results are consistent with the TRE evaluations using landmarks. Our method achieved the lowest RMSE compared to all the methods tested, with statistical significance ($p < 0.01$). The method also achieved the highest SSIM, but without statistical significance when compared against classic BP and Coop CAE.

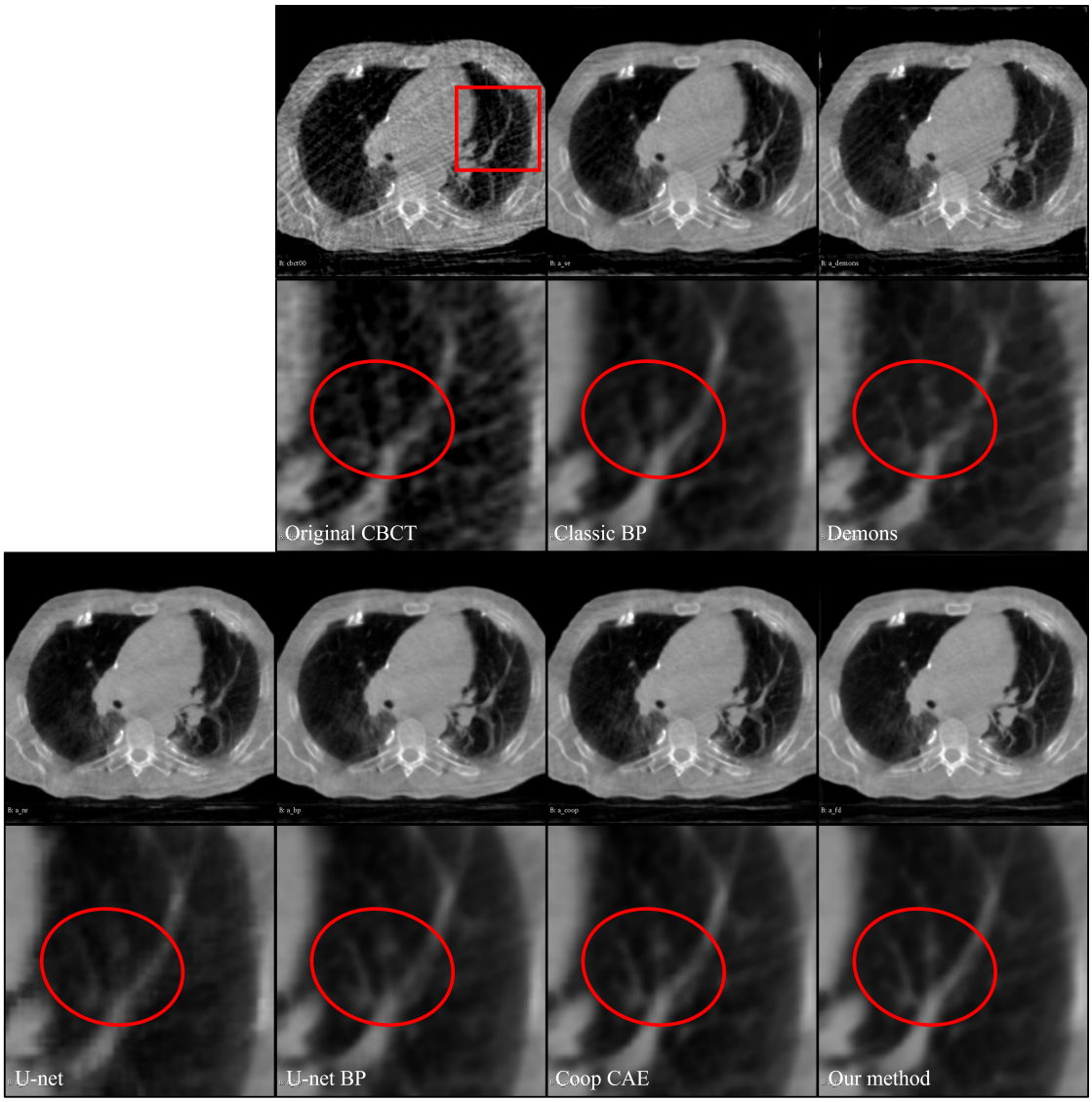

Figure 3.7: Example motion-compensated image enhancement results on real CBCT. Class BP, Demons still exhibit strong streak artifacts; U-net, U-net BP, Coop CAE, and our method show smoother images but our method has the sharpest detail.

3.2.3 Experiments on Cardiac CTA and MRI

3.2.3.1 Data

CTA dataset: The CTA dataset for DVF sample generation consists of 10 4D scans, each containing the ED and ES frames of a cardiac cycle. The scans used contrast according to

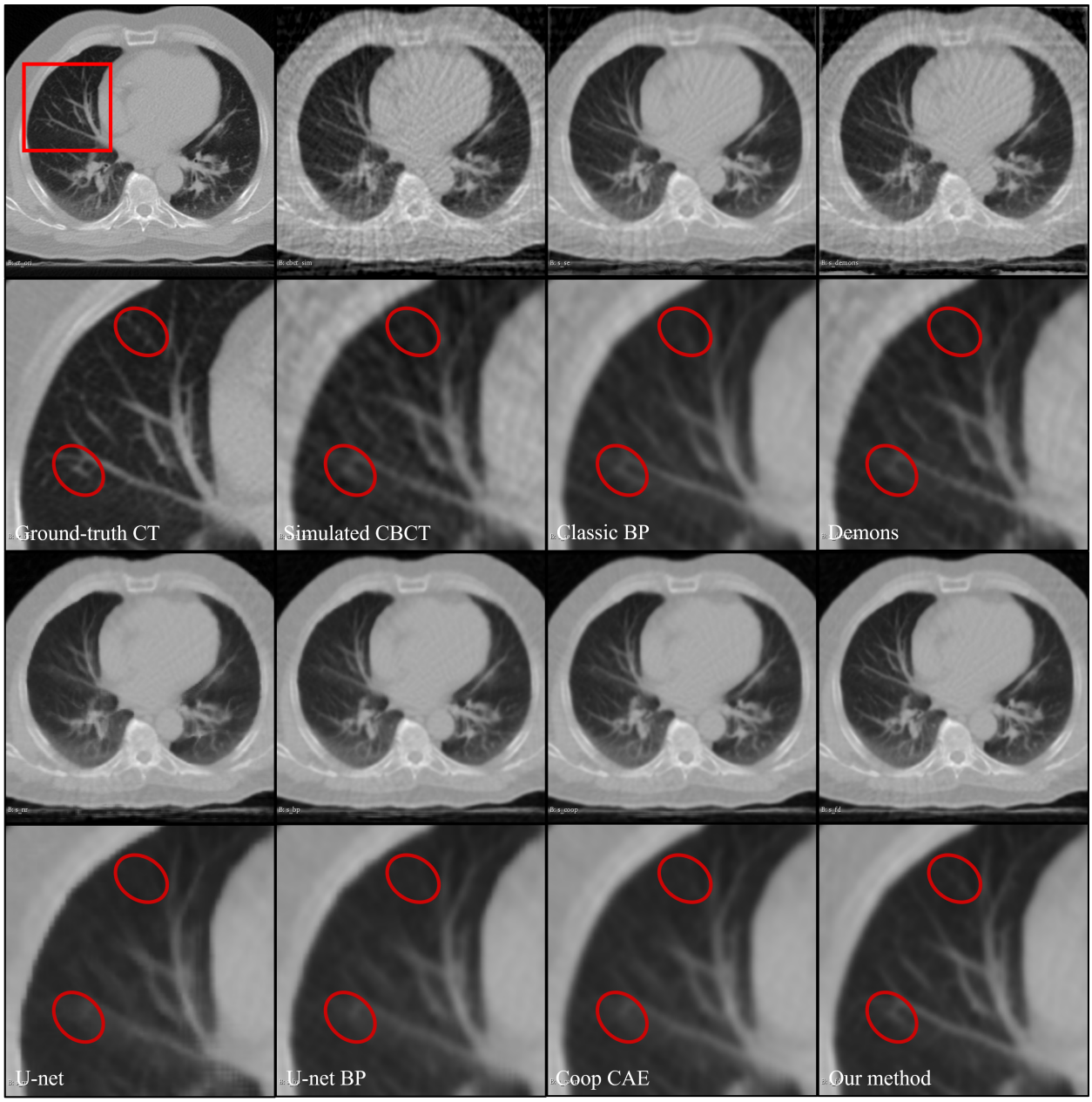

Figure 3.8: Example motion-compensated image enhancement results on simulated CBCT. The streak artifacts are better alleviated in the four deep learning methods. Sharper detailed structures are reconstructed in our method.

typical clinical system, on patients who are suspected to have cardiovascular problems (seven females, aged 74 $\pm$ 14 years). The image size was 512 $\times$ 512, with number of slices ranging

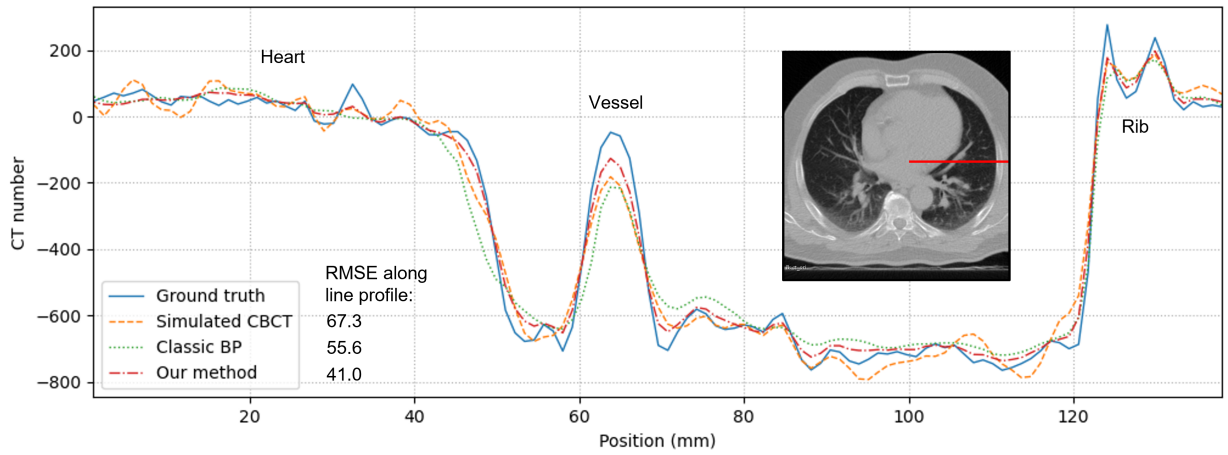

Figure 3.9: Profiles of the CBCT image before and after enhancement. The image intensity curves are from a horizontal line segment indicated on the image. The numbers indicate the root-mean-square errors to the ground truth for pixels on this line.

from 240 to 564, in-plane resolution ranging from 0.31 to 0.45 mm, and slice thickness ranging from 0.30 to 0.50 mm. The images were resampled with voxel spacing 0.5 mm and cropped with a $320 \times 320 \times 224$ window that covered the entire heart.

ViewRay 0.35T MRI dataset: The 0.35T MRI scans were acquired on a ViewRay MRIdian system with a balanced steady-state free precession (bSSFP) sequence, each containing the ED and ES frames. The scans were from patients who are suspected to have cardiovascular problems (five females, aged 55 ± 19 years) and were acquired in long-axis view with a breath-hold and EKG gating protocol. The slice thickness ranged from 6.00 to 8.00 mm with an average slice number of 25. The in-plane pixel spacing ranged from 1.25 to 2.18 mm. Each image was resampled with voxel spacing of 1 mm and cropped to $160 \times 160 \times 112$. The image intensity was cropped between 0 and 1000 and then normalized to [0,1].

We performed multi-compartment manual segmentation of the whole heart, left and right ventricles (LV & RV), left and right atria (LA & RA), and pulmonary artery (PA). An example ED-ES image pair and its segmentation are shown in Fig. 3.10.

Table 3.2: Motion-compensated CBCT enhancement results. Results are provided as mean ± standard deviation (p-value from paired t-tests).

	RMSE (HU)	SSIM
Original CBCT	150.2±11.85	0.982±0.004
Classic BP	115.3±7.83 (10^{-4})	0.991±0.003 (0.023)
Demons	117.2±8.69 (10^{-5})	0.987±0.003 (10^{-4})
U-net	123.3±8.72 (10^{-6})	0.988±0.002 (10^{-4})
U-net BP	118.5±8.11 (10^{-5})	0.990±0.003 (0.001)
Coop CAE	110.2±7.65 (0.003)	0.992±0.002 (0.052)
Our method	**107.5±8.51**	**0.993±0.002**

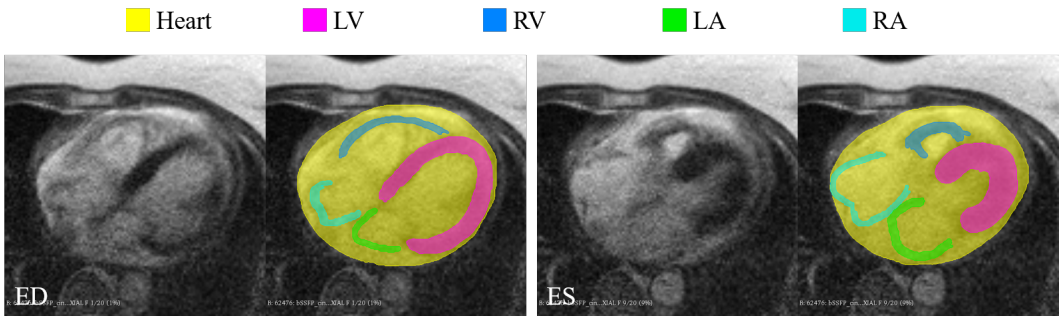

Figure 3.10: Example MRI image pair and the corresponding segmentation on an axial slice.

cMAC 3T MRI dataset: The 3T MRI scans were acquired as part of the cMAC public dataset [TDM13]. Each scan had 30 frames. The scans were from 15 healthy volunteers without clinical history of cardiac disease (three females, aged 28 ± 5 years) and were acquired in short-axis view with a breath-hold and EKG gating protocol. The image size was typically $256 \times 256 \times 14$, with voxel size $1.25 \times 1.25 \times 8$ mm^3 . The image was resampled to horizontal long-axis-view grid, with 1 mm voxel spacing. Then, they were cropped and normalized to the same size as the 0.35T images.

In each scan, 12 landmarks on LV were located using the corresponding tagged MRI:

one landmark per wall (anterior, lateral, posterior, septal) per ventricular level (basal, mid-ventricular, apical). The landmarks were manually tracked by two observers. The average inter-observer variability was 1.12 mm.

3.2.3.2 Training of the Feasibility Descriptor

When SimpleElastix was used to generate the DVF samples, all 10 CTA scans were registered 10 times, using different BP regularization weights, giving rise to 100 DVFs as the training set. With the deep learning methods, all 10 scans were processed with three different networks structures with two training settings each, giving rise to 60 DVFs as the training set. The diffeomorphic Demons method used 250 iterations and Gaussian smoothing with standard deviation of 0.5, 1.0 and 2.0, giving rise to 30 DVFs. An example CTA image and its derived SimpleElastix DVF are shown in Fig. 3.11.

With the three DVF sample sets, CAE based on SimpleElastix (CAE_{SE}), CAE based on deep learning methods (CAE_{DL}) and CAE based on all the methods (CAE_{All}) were trained independently, each with a batch size of 1 and 120,000 iterations. The ADAM optimizer with a learning rate of 10^{-4} was used.

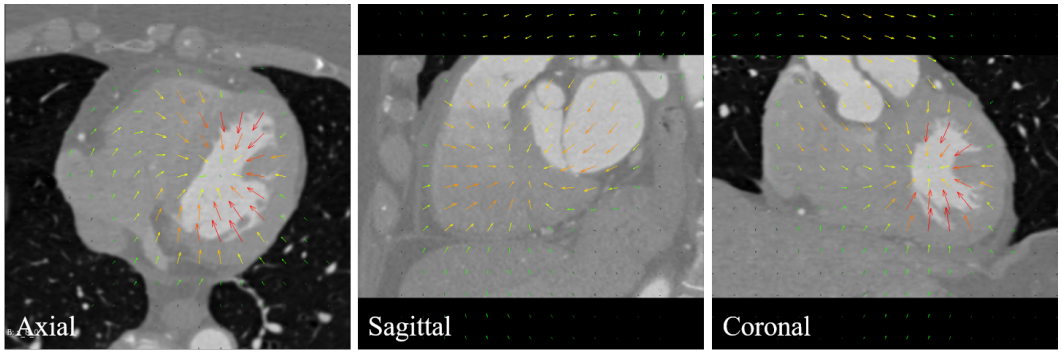

Figure 3.11: Example CTA image and DVF derived from classic B-spline registration.

3.2.3.3 Training of the DVF Estimation Network and Assessment

0.35T images and segmentation-based performance evaluation We performed a leave-one-out cross-validation experiment on the 0.35T scans: each time nine scans were used for training and the remaining one for testing. The network was trained for 2000 epochs with a batch size of one. The balancing parameter μ was tuned based on the training performance, and was set to 10^{-7} subsequently. The ADAM optimizer with a learning rate of 10^{-4} was used. For comparison, U-nets with and without BP were tuned, trained, and tested with the same setting. The weight for BP was tuned to be 0.1.

Evaluation metrics include the Dice similarity coefficient (DSC), Hausdorff distance (HD) and 80% Hausdorff distance ($HD_{80\%}$) between the fixed mask and estimated mask propagated from the moving instance. Wilcoxon signed-rank test was used to examine the performance improvement.

3T images and landmark-based performance evaluation From the cMAC dataset, 15 scans were split into a set of five for refinement training and a set of ten for hold-off test and performance evaluation. The refinement training was performed to fit the model to the specific imaging protocol in the dataset. The pre-trained model on 0.35T images was used for initialization and the refinement training took 20 epochs. Other hyperparameter settings were kept the same as 3.2.3.3.

Euclidean distance between the fixed and the transformed landmarks was calculated as TRE to measure registration performance. Paired t-test was performed to examine the performance improvement.

3.2.3.4 Results

To illustrate the performance of the CAE-based regularizer, an example input DVF for the CAE and its corresponding DVF reconstructed from the CAE-output Jacobian are shown in Fig. 3.12. The output DVF was computed using the boundary conditions acquired from

the input DVF and reconstructed from the Jacobian by integration. It can be observed that the reconstructed output preserved the motion pattern but was smoother, as the DVF complexity was constrained by the latent space. Note that the input DVF is an intermediate result from an incomplete training process, and that the reconstructed DVF is calculated from the output Jacobian by integration, which is not required as a step in our method, but is illustrated here to demonstrate the effectiveness of using CAE as a regularizer.

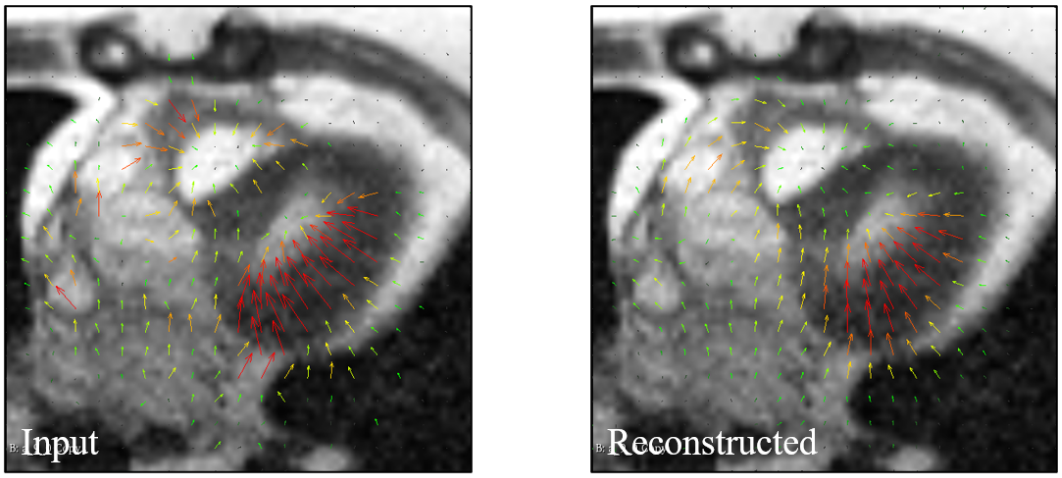

Figure 3.12: Example CAE input DVF and its reconstructed output. The background is the corresponding fixed image.

Fig. 3.13 and Fig. 3.14 show some example registration results on the ViewRay and cMAC data, respectively. Despite some differences in local deformation magnitude, all the three CAE-generating methods resulted in reasonable and smooth DVFs with good intensity matching. U-net without regularization generated non-smooth and non-feasible DVF. U-net BP regularization achieved both good intensity matching and smooth DVF, but the DVF was not physically sound with the large tangential component of the deformation vectors near the left ventricle, as indicated by the red circle. In comparison, the DVF generated from our method was physically more feasible.

We have typically observed larger registration errors in the RV and RA regions, regardless of imaging protocol or patient diseases. The errors were mostly caused by strong image

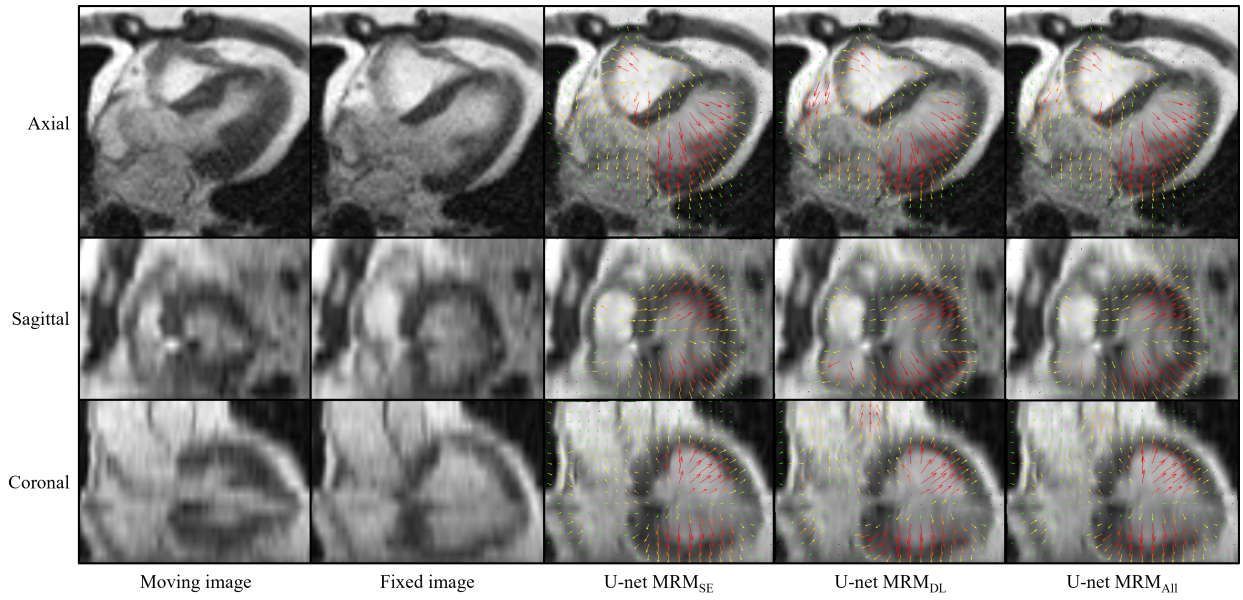

Figure 3.13: Example registration results on ViewRay data.

artifacts. The thin wall structures of RV and RA myocardium made the registration especially challenging. Fig. 3.15 shows an example where our method failed to estimate the RV motion correctly. In this case, the image intensity gradient was too strong for the motion prior to overcome, and our method behaved similarly to the BP regularization.

Table. 3.3 shows the quantitative results based on segmentation on the ViewRay data. In terms of $HD_{80\%}$, CAE_{SE} achieved the best result on RV, and CAE_{All} achieved the best result on LV and PA, with statistically significant improvement over U-nets with and without BP. On LA and RA, our method was comparable to the best results without appreciable statistical significance. The HD results were consistent with the $HD_{80\%}$ in general, with the three CAE methods achieving the best results on LV, RV, and PA. The error reduction (comparing HD before and after registration) was less obvious due to the quality of the ground truths and the sensitivity to artificial outliers. Across the metrics, our method achieved comparable results to the best one of the two settings in B-spline registration, and the three CAE-generating methods yielded similar accuracy.

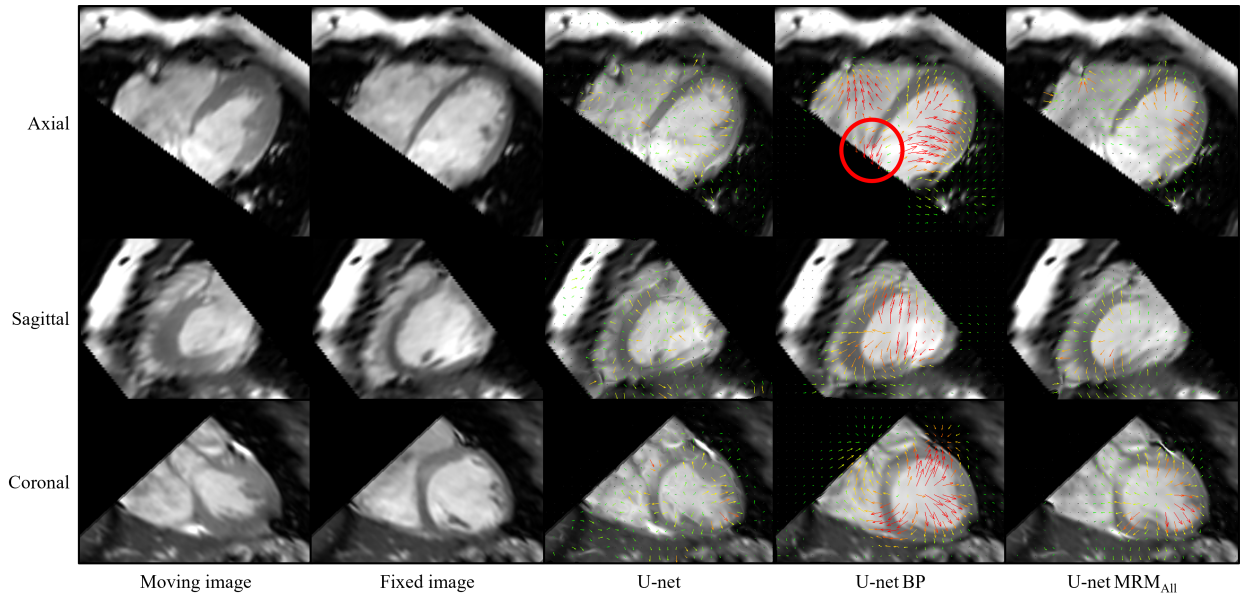

Figure 3.14: Example registration results on cMAC data. The red circle indicates unrealistic motion tangential to the ventricle.

Table. 3.4 shows the quantitative results based on landmarks on the cMAC test. The average Euclidean distance between the same tagging location annotated by the two observers was calculated as inter-observer variability. The variabilities for the basal, midventricular, and apical regions were 1.17, 1.05, 1.14 mm respectively. The three CAE-generating methods resulted in similar TREs to each other, without statistical significance, indicating that our method was robust against different DVF sample generation methods. Our method achieved the lowest TREs in the basal and midventricular regions, and was close to the best-performing B-spline registration in the apical region. When considering all the landmarks, our method achieved the best TRE, significantly lower than the other methods tested.

The average registration time was 40 s for the classic B-spline registration and 0.02 s for all the networks.

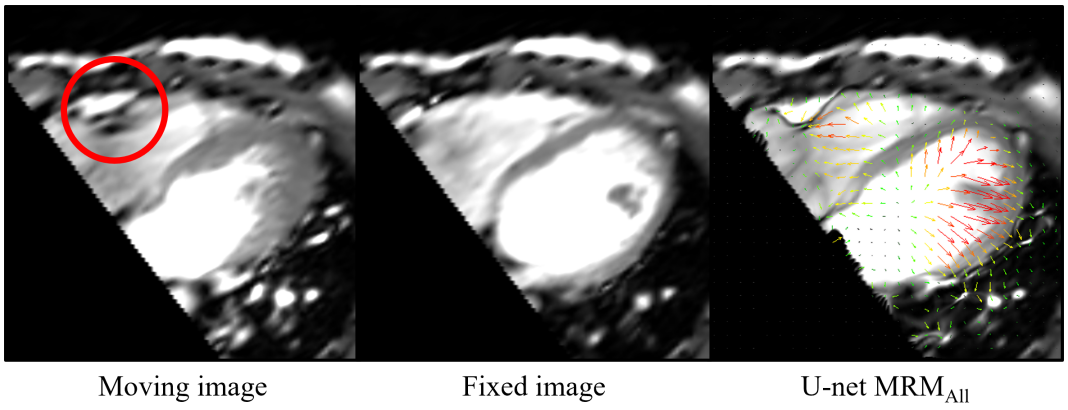

Figure 3.15: An example case where our method failed to estimate the motion correctly due to the strong artifact (red circle).

3.3 Unupervised Feasibility Prior Based on Statistical Generative Model

3.3.1 Method

3.3.1.1 The Overall Image Registration Framework

As shown in Fig. 3.16, the proposed registration framework consists of a generator network and a spatial transformation module. The generator network is trained in an unsupervised fashion using fixed and moving image pairs to implicitly model the DVF feasible set (details in Sec. 3.3.1.3). The trained generative model imposes constraints on deformation by generating DVFs from low dimensional latent variables (vector). This new parametrization serves as a more powerful alternative to the explicit parametrizations as in the conventional B-spline model. It can encode spatially heterogeneous properties of physical or physiological motion beyond smoothness offered by pre-defined fixed order polynomial basis.

The overall function can be written as:

$$x' = F(x, z; \hat{\theta}) = F_T(x, F_G(z; \hat{\theta})), \tag{3.3}$$

Table 3.3: Assessment of agreement between structure delineation between warped and fixed images. Results are provided as mean ± standard deviation. Numbers inside the parentheses indicate p-value results from Wilcoxon signed-rank tests when comparing to CAE_{All}. p-values that indicate statistical significance (< 0.01) are underlined. Values in bold indicate the best results.

	DSC	$\text{HD}_{80\%}$ (mm)					HD (mm)				
———	Heart	LV	RV	PA	LA	RA	LV	RV	PA	LA	RA
Before	0.928±0.012 (0.83)	4.94±1.42 ($\underline{10^{-6}}$)	10.34±4.28 ($\underline{10^{-5}}$)	5.40±1.91 ($\underline{10^{-4}}$)	6.46±2.02 ($\underline{10^{-3}}$)	10.34±5.84 ($\underline{10^{-4}}$)	8.64±5.31 ($\underline{10^{-6}}$)	15.85±8.99 ($\underline{10^{-3}}$)	9.16±4.70 ($\underline{10^{-3}}$)	10.64±5.65 ($\underline{10^{-3}}$)	15.53±9.26 ($\underline{10^{-3}}$)
B-spline 0.1	0.922±0.022 ($\underline{10^{-5}}$)	2.37±0.90 (0.06)	7.41±4.72 (0.63)	4.84±2.50 (0.18)	5.39±3.08 (0.70)	**8.56±6.05** (0.28)	5.20±3.31 (0.38)	14.41±6.27 (0.45)	8.43±4.30 ($\underline{10^{-3}}$)	9.81±4.59 (0.49)	14.75±7.30 (0.72)
B-spline 1	0.927±0.017 (0.95)	2.27±0.69 (0.13)	7.20±3.99 (0.95)	4.52±2.32 (0.87)	5.62±3.10 (0.75)	8.78±5.92 (0.92)	5.11±3.42 (0.88)	14.29±6.03 (0.81)	7.74±3.93 (0.31)	**9.63±5.01** (0.53)	**14.55±7.65** (0.75)
U-net	0.921±0.017 ($\underline{10^{-6}}$)	2.99±1.36 ($\underline{10^{-3}}$)	8.80±4.31 (0.03)	5.56±1.87 ($\underline{10^{-3}}$)	6.28±2.92 ($\underline{10^{-3}}$)	9.51±5.93 ($\underline{10^{-3}}$)	6.59±3.99 ($\underline{10^{-4}}$)	15.59±7.34 ($\underline{10^{-3}}$)	9.40±4.93 ($\underline{10^{-4}}$)	10.91±5.31 ($\underline{10^{-3}}$)	15.41±8.22 ($\underline{10^{-3}}$)
U-net BP	**0.928±0.019** (0.08)	2.45±1.10 ($\underline{0.01}$)	7.88±3.94 (0.14)	4.61±1.89 (0.29)	**5.09±2.47** (0.11)	8.88±6.29 (0.33)	5.40±4.12 ($\underline{0.01}$)	15.08±6.59 (0.43)	8.02±3.89 ($\underline{0.01}$)	9.67±4.43 (0.23)	14.88±7.75 (0.49)
U-net CAE_{SE}	0.927±0.018 (0.90)	2.25±0.72 (0.49)	**7.20±3.54** (0.72)	4.43±1.93 (0.99)	5.49±2.58 (0.58)	8.86±5.92 (0.81)	**5.10±3.92** (0.87)	**14.18±6.03** (0.71)	7.52±4.08 (0.97)	9.88±4.86 (0.55)	14.84±7.86 (0.73)
U-net CAE_{DL}	0.926±0.022 (0.88)	2.23±0.87 (0.37)	7.24±4.02 (0.31)	4.58±2.23 (0.35)	5.39±2.91 (0.95)	8.75±6.10 (0.90)	5.17±3.77 (0.61)	14.55±6.44 (0.70)	7.55±4.17 (0.49)	9.92±4.95 (0.55)	14.77±7.76 (0.90)
U-net CAE_{All}	0.927±0.021	**2.23±0.77**	7.21±3.79	**4.42±1.96**	5.39±2.72	8.71±5.62	5.11±4.01	14.22±6.30	**7.37±3.88**	9.91±4.99	14.87±7.87

where $x \in \mathbb{R}^{d_1 \times d_2}$ is the input moving image, x' is the warped image, $z \in \mathbb{R}^{d_z}$ is the latent deformation variables, $\hat{\theta}$ is the trained network parameters, F_G is the function of the generator network, and F_T is the spatial transformation.

When performing registration, gradient descent is used for optimization with respect to the latent deformation variables z, rather than the dense DVF. Optimization is driven by similarity between the fixed image y and the warped moving image x', without extra regularization. Given a moving and fixed image pair (x, y), the optimal DVF $\hat{v}$ can be achieved by:

$$\hat{z} = \underset{z}{\operatorname{argmin}}\, L_{sim}(y, x') = \underset{z}{\operatorname{argmin}}\, L_{sim}(y, F_T(x, F_G(z; \hat{\theta}))), \tag{3.4}$$

$$\hat{v} = F_G(\hat{z}; \hat{\theta}). \tag{3.5}$$

Table 3.4: Target registration errors based on landmarks. Results are provided as mean ± standard deviation in millimeter. Numbers inside the parentheses indicate p-value results from paired t-tests when comparing to CAE_{All}. p-values that indicate statistical significance (< 0.01) are underlined. Values in bold indicate the best results.

	Basal	Midventricular	Apical	All
Before	8.73±2.20 ($\underline{10^{-22}}$)	7.82±2.90 ($\underline{10^{-16}}$)	7.30±3.24 ($\underline{10^{-16}}$)	7.95±2.88 ($\underline{10^{-50}}$)
B-spline 0.1	2.41±1.24 ($\underline{10^{-3}}$)	2.62±1.43 (0.51)	**1.78±1.18** (0.81)	2.27±1.31 (0.09)
B-spline 1	2.37±1.33 (0.08)	2.66±1.48 (0.37)	1.85±1.25 (0.17)	2.30±1.35 ($\underline{10^{-3}}$)
U-net	2.98±2.01 ($\underline{10^{-4}}$)	3.22±2.35 ($\underline{10^{-5}}$)	2.20±1.66 ($\underline{10^{-4}}$)	2.80±2.28 ($\underline{10^{-5}}$)
U-net BP	2.38±1.29 (0.16)	2.71±1.45 (0.03)	1.92±1.31 ($\underline{0.01}$)	2.34±1.40 ($\underline{10^{-3}}$)
U-net CAE_{SE}	2.31±1.54 (0.92)	2.60±1.42 (0.67)	1.82±1.18 (0.60)	2.25±1.49 (0.90)
U-net CAE_{DL}	**2.28±1.45** (0.87)	2.62±1.48 (0.86)	1.79±1.15 (0.57)	2.23±1.46 (0.88)
U-net CAE_{All}	2.30±1.45	**2.59±1.48**	1.82±1.27	**2.23±1.44**

In this work, we use NCC as the image similarity metric, but the method applies to other choices such as mutual information [VVS20].

3.3.1.2 Architecture of the Statistical Generator Network

The model was implemented in both 2D and 3D. The 2D version of the generator network is shown in Fig. 3.17. The network takes latent variables z as input and outputs a DVF. It first applies a fully connected layer to map the latent variables to image space. Then, two 3×3 deconvolution layers and two 5×5 deconvolution layers are applied. A residual block [HZR16] containing two 3×3 convolutional layers is added between the two 5×5 deconvolution layers, which encourages connections between neighboring "patches", while preventing the training from vanishing gradients at the same time. The convolutional layers use a stride of 1. All the deconvolution layers use a stride of 2. All the layers use zero-padding and ReLU activation, except that the last layer uses tanh activation. After tanh,

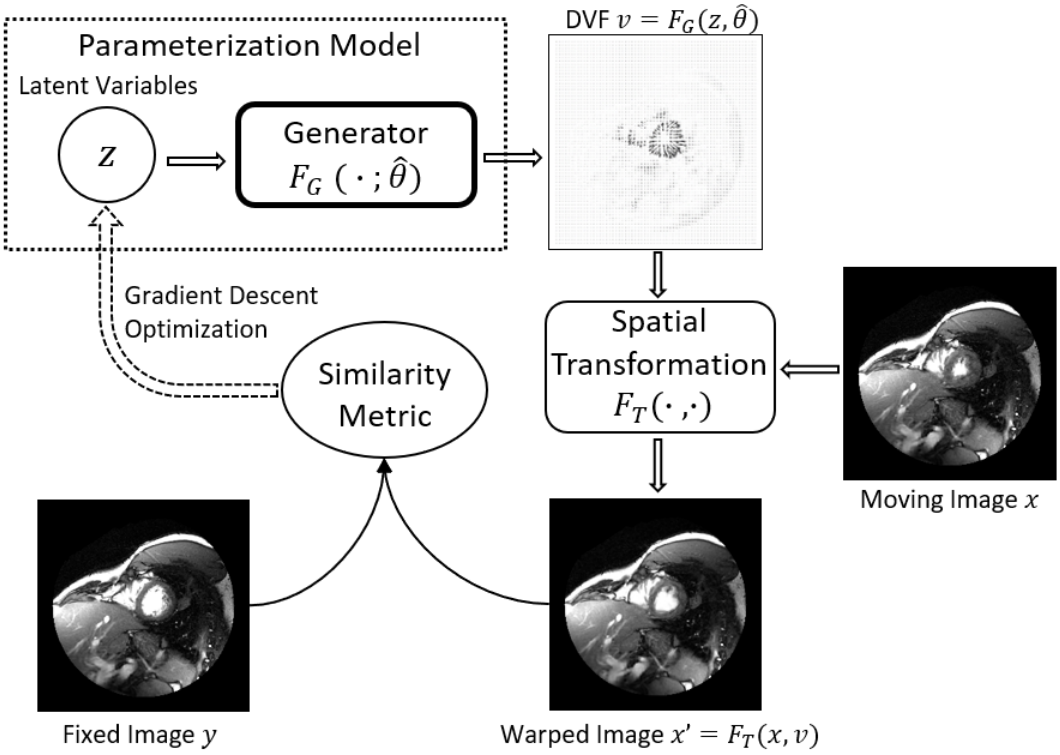

Figure 3.16: Overview of the proposed generator-based registration framework. The trained generator network takes latent deformation variables z as input and outputs a DVF. Then the spatial transformation module warps the moving image towards the fixed with this DVF. During registration, image similarity is calculated, which drives the gradient descent with respect to z to achieve the optimal DVF for each image pair.

the DVF is scaled to pixel unit before fed into the spatial transformation. The generator network for 3D DVF takes a similar form, with the same numbers of channels and layers, and (de)convolutions replaced by 3D counterparts.

3.3.1.3 Alternating Back-propagation for Generator Optimization

To learn this generative model, we employ a maximum-likelihood learning and inference algorithm called alternating back-propagation [HLZ17]. Training of the model requires both

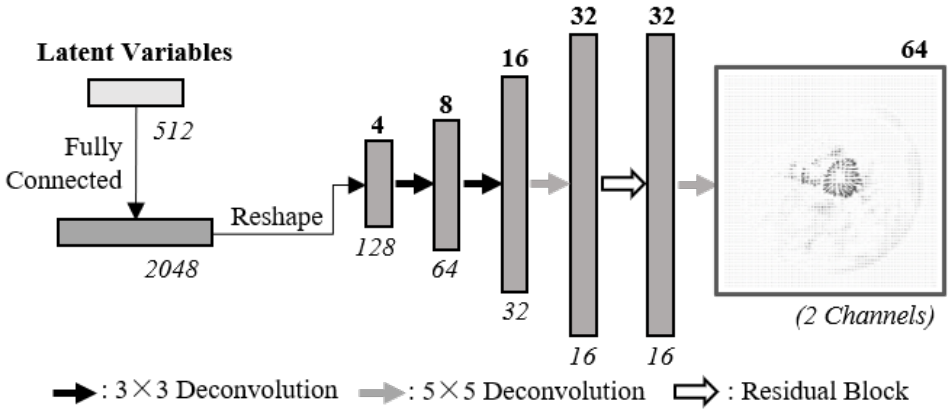

Figure 3.17: Architecture of the generator network (2D version). Numbers above and below the blocks indicate tensor sizes and numbers of channels, respectively.

inferential back-propagation on latent variables z_i and learning back-propagation on network parameter θ, to maximizing the log-likelihood on the training image pairs $\{(x_i, y_i), i = 1, ..., N\}$:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log p(y_i|x_i;\theta) = \frac{1}{N}\sum_{i=1}^{N} \log \int p(y_i, z_i|x_i;\theta)dz_i. \tag{3.6}$$

The gradient of $L(\theta)$ can be calculated according to the expectation maximization algorithm:

$$L'(\theta) = \frac{\partial}{\partial\theta}\log p(y|x;\theta) = \frac{1}{p(y|x;\theta)}\frac{\partial}{\partial\theta}\int p(y, z|x;\theta)dz = \mathbb{E}\left[\frac{\partial}{\partial\theta}p(y, z|x;\theta)\right]. \tag{3.7}$$

Langevin dynamics is employed to obtain an approximation to the expectation, by sampling the posterior $p(z|x, y;\theta)$ and then computing the Monte Carlo average. The latent variables are updated in the inferential back-propagation:

$$z^{s+1} = z^s + \frac{\delta^2}{2}\frac{\partial}{\partial z}\log p(y, z^s|x;\theta) + \delta\varepsilon_s, \tag{3.8}$$

where s in the superscript is the index of Langevin sampling steps, δ is the step size and $\varepsilon \sim \mathrm{N}(0, I_{d_z})$ is a random vector. The log of the joint density can be evaluated by:

$$\log p(y, z|x;\theta) = \log\left[p(z)p(y|z, x;\theta)\right] = -\frac{1}{2}\left\|z\right\|^2 - f(y, F(x, z;\theta)) + \text{const.}, \tag{3.9}$$

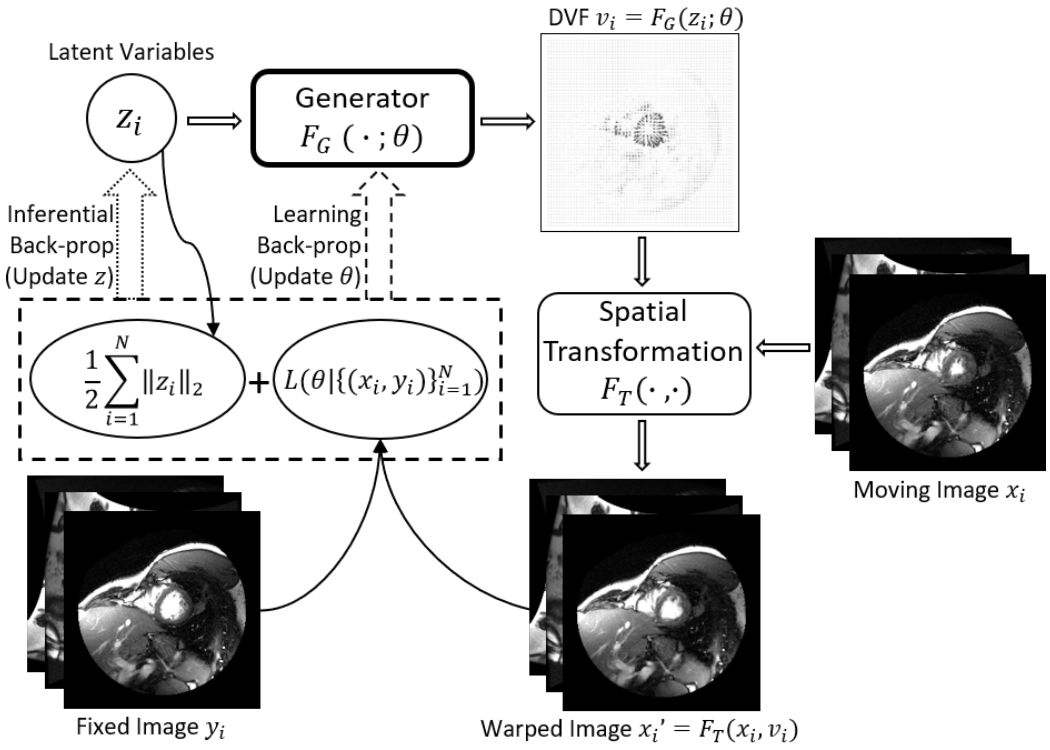

Figure 3.18: Training of the generative model. The generator network is trained in an unsupervised setting by alternatively updating z through the inferential back-propagation and θ through the learning back-propagation to maximize the likelihood on the training set $\{(x_i, y_i), i = 1, ..., N\}$.

where $f(y, F(x, z; \theta))$ corresponds to L2 dissimilarity $f(y, x') = \frac{1}{2\sigma^2}\|y - x'\|^2$ when we further assume $y = F(x, z; \theta) + \epsilon$, $\epsilon \sim \mathrm{N}(0, \sigma^2 I)$. Here, we adopt a slightly more robust variation and let

$$f(y, x') = \frac{1}{2\sigma^2}(1 - \mathrm{NCC}(y, x')), \tag{3.10}$$

assuming that y follows an "exp-NCC" distribution.

Given a training sample (x_i, y_i), we first apply the Langevin dynamics in Eq.(3.8) to get the corresponding latent variables z_i, and then use it to compute $L'(\theta)$ in Eq.(3.7) through

Monte Carlo approximation:

$$L'(\theta) \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \theta} \log p(y_i, z_i|x_i; \theta) = -\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \theta} f(y_i, F(x_i, z_i; \theta)). \tag{3.11}$$

The gradient is then used for updating θ in the learning back-propagation: $\theta_{t+1} = \theta_t + \eta L'(\theta_t)$, where t is the index of training iterations and η is the learning rate.

In the algorithm, we alternate between the inferential and learning back-propagations to jointly optimize the network and the latent variables for the training set. To reduce computational cost, warm start is used: at the beginning of each inferential loop, the latent variables z_i are initialized from the previous training iteration. Fig. 3.18 and Algorithm. 1 describe the details of the algorithm.

Algorithm 1 Alternating back-propagation

Input:

(1) Training image pairs $\{(x_i, y_i), i = 1, ..., N\}$

(2) Number of Langevin steps S

(3) Number of learning iterations T

Output:

(1) Learned network parameters θ

(2) Inferred latent variables for the training set $\{z_i, i = 1, ..., N\}$

1: Let $t \leftarrow 0$, initialize θ.

2: Initialize $\{z_i, i = 1, ..., N\}$ from Gaussian distribution.

3: **while** $t \leq T$ **do**

4: **Inferential back-propagation:** For each i, run S steps of Langevin dynamics to sample z_i from $p(z_i|x_i, y_i; \theta)$. Starting from the current z_i, each step follows Eq. (3.8).

5: **Learning back-propagation:** Update the network parameters $\theta_{t+1} \leftarrow \theta_t + \eta L'(\theta_t)$, with learning rate η, where $L'(\theta_t)$ is computed according to Eq. (3.11).

6: Let $t \leftarrow t + 1$.

7: **end while**

3.3.2 Experiments on 3D Synthetic Images

3.3.2.1 Data Generation

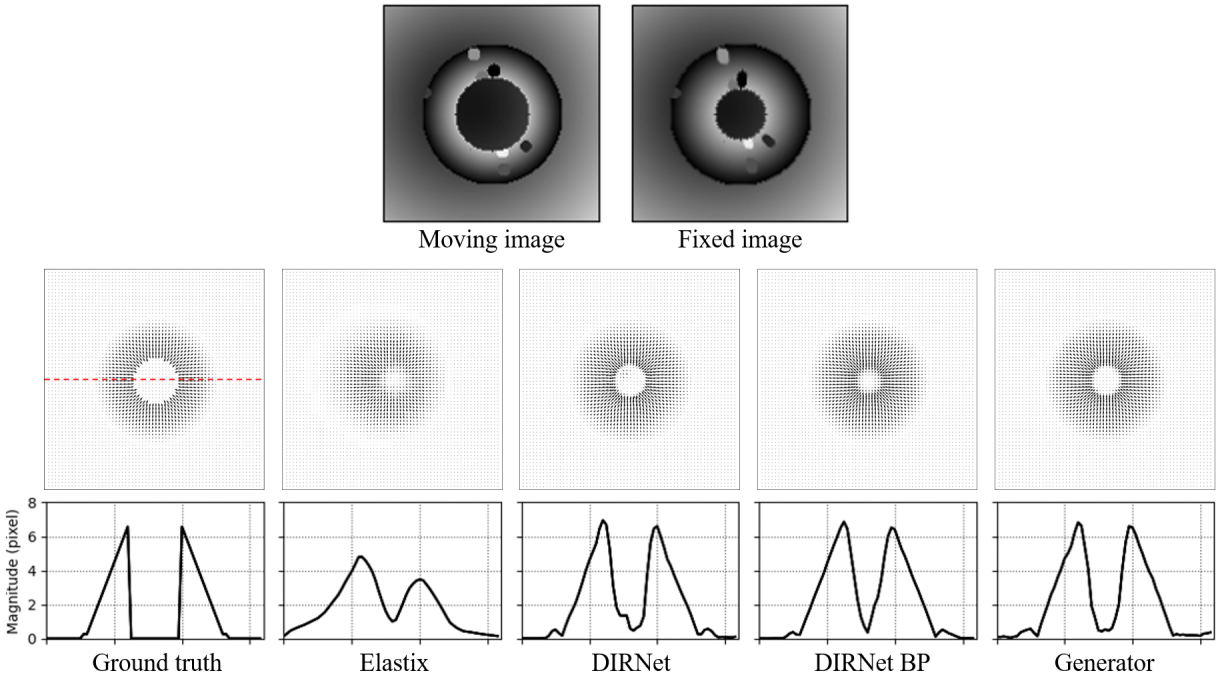

Figure 3.19: Example registration results on synthetic images. One slice from a 3D volume is shown. The bottom row shows DVF profiles for the location indicated by the red dashed line.

Given the absence of ground-truth DVFs in clinical 3D images, we generated digital phantoms for a quantitative evaluation on dense DVFs. An example synthesized image pair (one slice from a 3D volume) is shown in Fig. 3.19 (a-c). First, we generated moving shell objects as foregrounds in moving images. The positions and inner and outer radii of the spherical shell were randomly selected. The intensity values inside the spheres decreased as the distance from the spherical center increases. Small spheres with random locations and grey values were superimposed on the foreground to simulate local textures. DVFs were simulated to radially shrink the moving shells. The DVF magnitude increased linearly as it moves closer to the spherical center. The maximum DVF magnitude was randomly

selected. Fixed images were generated by applying the DVFs to the moving images. The image resolution was $128 \times 128 \times 128$. Training, validation, and testing sets used 4096, 256, and 256 synthetic samples, respectively.

3.3.2.2 Model Training

The length of latent variables d_z was reduced to 256 considering the low complexity of deformation in this experiment. The parameter σ controlling the strength of NCC in Eq. (3.10) was set to 0.002. The model was trained in mini-batches of 4 samples for 1500 epochs. The inferential back-propagation used 10 Langevin steps with the step size δ set to 0.01. The learning back-propagation used ADAM optimizer [KB14] with the learning rate η set to 10^{-4}.

3.3.2.3 Evaluation

Root mean squared errors (RMSEs) to the ground-truth DVFs were calculated as a dense version of target registration error (TRE_d) to indicate registration performance:

$$TRE_d = \sqrt{\frac{1}{d_1 d_2 d_3} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \sum_{k=1}^{d_3} \|v^*_{ijk} - v_{ijk}\|^2}, \tag{3.12}$$

where v^* and v are the ground-truth and result DVFs.

3.3.2.4 Result

Fig. 3.19 shows an example result of the proposed method in comparison to SimpleElastix (with the weight λ for bending energy carefully tuned for each specific case to achieve the best possible trade-off) and DIRNet trained with and without using bending energy penalty (BP) regularization [VBV19]. From the DVF magnitude profiles in Fig. 3.19 (h), it can be observed that our method was more robust to local textures and the motion boundary was closer to the ground truth, with motion discontinuity on the inner surface. The large

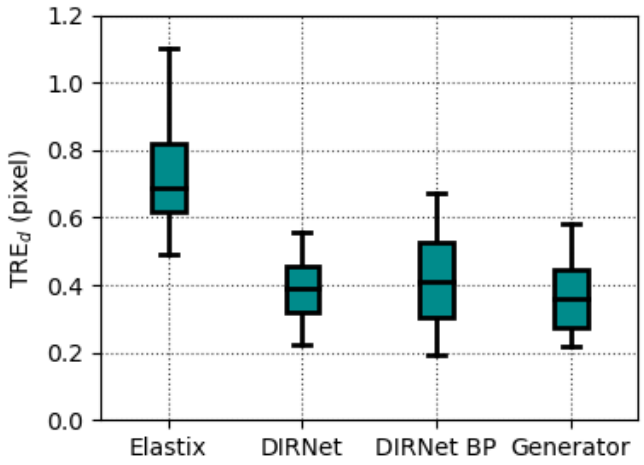

Figure 3.20: Boxplots showing the synthetic image registration results.

($\lambda = 40$) bending energy in SimpleElastix enlarged the area of deformation. Introducing bending energy to the DIRNet removed the large false estimation in the interior background region surrounded by the shell object, but compromised the definition of motion boundary. Paired t-tests on RMSEs show that our method reduced registration error significantly, over other methods tested, as shown in Table. 3.5 and Fig. 3.20.

Table 3.5: Results of the synthetic image registration experiment. Registration errors are provided as mean ± standard deviation and the p-value from paired t-test against the result in our method.

	Elastix	DIRNet	DIRNet BP	Generator
TRE$_d$ (pixel)	0.73±0.14	0.39±0.09	0.42±0.13	**0.37±0.10**
p-value	2.89×10^{-70}	8.65×10^{-18}	6.92×10^{-26}	–

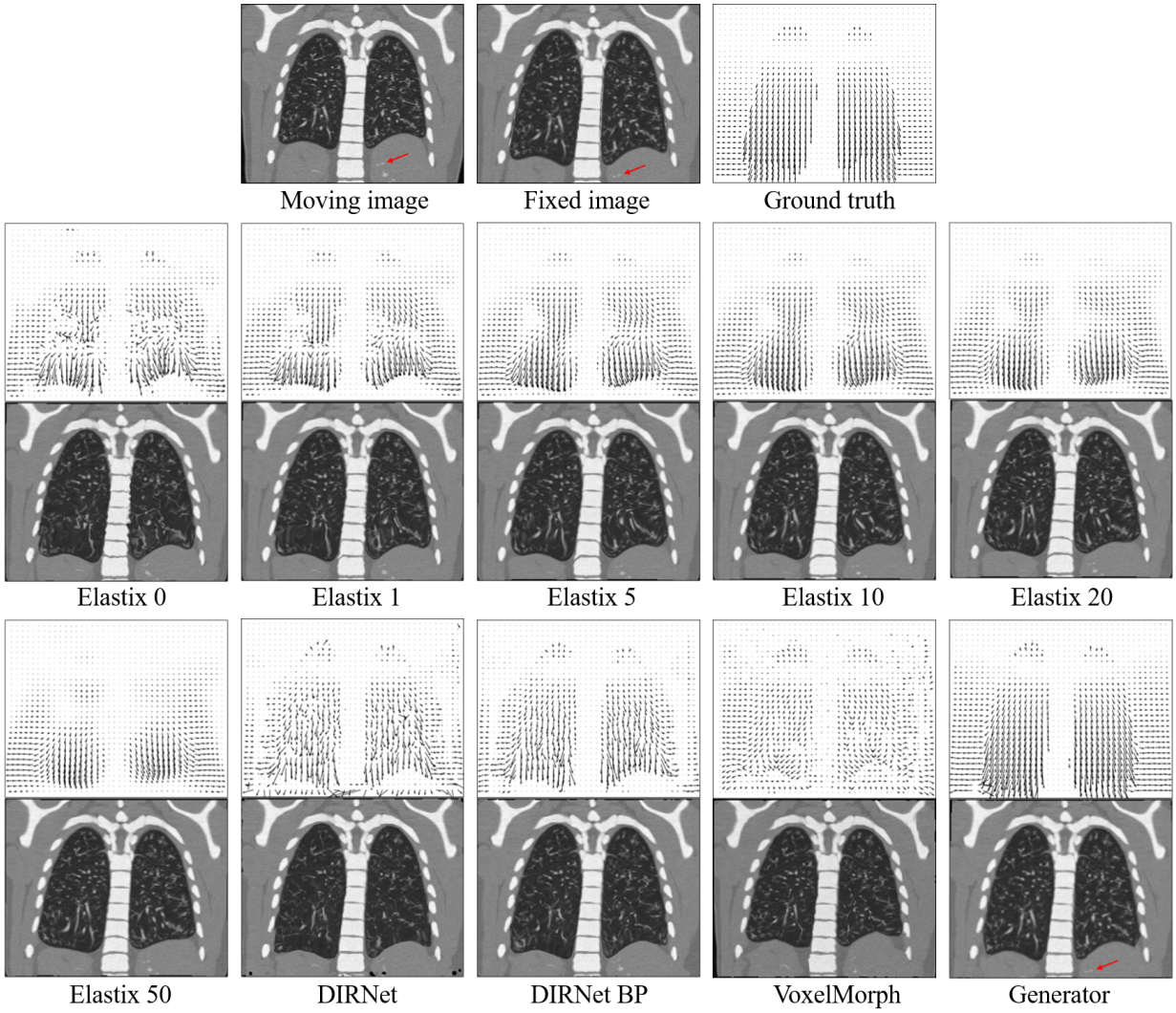

Figure 3.21: Example registration results on simulated CT images. Numbers for Elastix registrations indicate the weight λ for bending energy. Red arrows indicate local texture in the subdiaphragmatic region.

3.3.3 Experiments on Simulated CT

3.3.3.1 Data Generation

The simulated CT images were generated with the XCAT anthropomorphic digital phantom [SSM10, ASS17]. First, 10 slices of a simulated CT scan were selected as moving images, with

image resolution 256×320. Then using the segmentation mask of the lung, corresponding 2D DVFs were generated to simulate respiratory motion. An example simulated image pair is shown in Fig. 3.21 (a-c). The parameters for motion magnitude and zero-motion position were uniformly selected within a range. 130 DVFs were generated for each moving image. These DVFs were used to generate the warped images in the forward direction. 1300 synthetic samples were divided into training, validation, and testing sets, containing 1024, 176, and 100 samples, respectively.

3.3.3.2 Model Training

The model was trained in mini-batches of 8 samples for 2500 epochs, with other hyper-parameters same to Sec. 3.3.2.2.

In addition, to test the model's sensitivity to the latent variables length d_z, the model was also trained with 10 d_z settings ranging from 32 to 320.

3.3.3.3 Evaluation

RMSE to the ground-truth DVFs were calculated as a dense version of target registration error (TRE_d). Target registration errors on anatomical landmarks (TRE_l) were calculated using the landmarks at the lung and ribcage regions. Five landmarks were annotated for each image pair in the testing set. They were identified in the moving images first, and then their positions in the fixed images were calculated using the ground-truth DVFs.

3.3.3.4 Result

Fig. 3.21 shows an example result of the proposed method in comparison to SimpleElastix with a wide range of regularization weights λ between 0 and 50, DIRNet, and VoxelMorph. The single universal weight for bending energy in both SimpleElastix and DIRNet BP did not achieve a good balance between the robustness to local textures and heterogeneity for

different tissues. With moderate regularization ($\lambda < 5$), SimpleElastix generated DVFs that were non-smooth and non-diffeomorphic. As more prominent regularization was applied ($\lambda > 10$), the DVF became smoother but the image agreement was compromised severely, resulting in significant underestimation of DVF magnitude. The discontinuity along the lung surface made the diffeomorphic parametrization in VoxelMorph less suitable for this task, resulting in a blurred edge and an underestimated deformation magnitude. The simulated expansion of the lung was better characterized and the discontinuity along the lung contour was better captured in our method. In addition, our method also showed significantly superior motion reconstruction in the subdiaphragmatic regions, while all the other competing methods suffered from the lack of local image texture.

Table. 3.6 and Fig. 3.22 show the quantitative results. Our method outperformed the best manually tuned SimpleElastix, DIRNet, and VoxelMorph, with paired t-tests on TREs showing significantly lower registration errors based on both dense DVFs and landmarks.

Table 3.6: Results of the simulated CT registration experiment. Registration errors are provided as mean ± standard deviation and the p-value from paired t-test against the result in our method.

	Elastix0	Elastix1	Elastix5	Elastix10	Elastix20	Elastix50	DIRNet	DIRNet BP	VoxelMorph	Generator
TRE_d (mm)	3.85±1.57	3.44±1.44	3.14±1.27	3.04±1.19	3.10±1.23	3.15±1.30	5.15±0.91	3.48±1.09	4.39±1.53	**2.56±0.56**
p-value	4.71×10^{-18}	4.32×10^{-12}	3.35×10^{-8}	7.35×10^{-7}	1.02×10^{-7}	4.72×10^{-8}	8.24×10^{-71}	1.69×10^{-24}	1.43×10^{-29}	–
TRE_l (mm)	2.70±3.21	2.29±2.58	1.88±1.89	1.82±1.75	1.96±1.92	2.01±1.96	2.41±3.45	2.19±3.25	3.33±3.90	**0.96±1.20**
p-value	2.80×10^{-45}	1.52×10^{-41}	1.77×10^{-37}	4.44×10^{-39}	1.64×10^{-42}	2.41×10^{-43}	2.12×10^{-30}	2.66×10^{-24}	1.07×10^{-43}	–

Fig. 3.23 shows the results of the sensitivity test on d_z. It can be seen that both the loss value and the TRE were kept low for $d_z \geq 160$. The model required more training epochs to fully converge with larger d_z, therefore there was a slight increase of the loss and TRE when fixing the number of epochs to 2500. Overall, the model is not sensitive to d_z and does not require a trail and tuning process once d_z is roughly set based on DVF size and complexity.

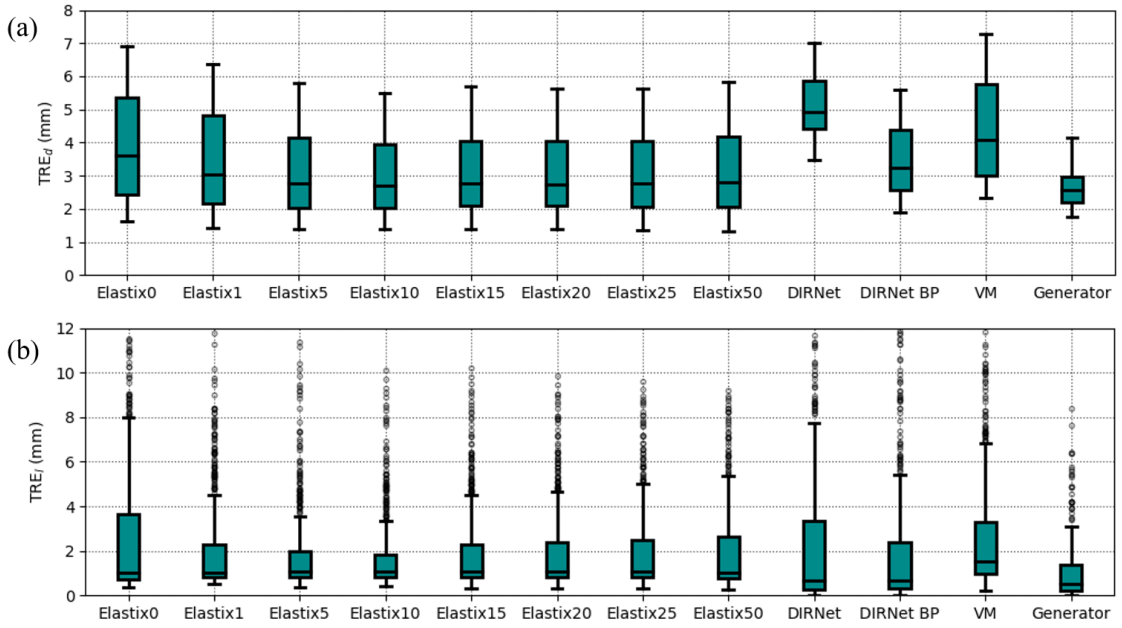

Figure 3.22: Boxplots showing the simulated CT registration results. (a) Target registration error on dense DVFs. (b) Target registration error on anatomical landmarks.

3.3.4 Experiments on 2D Cardiac MRI

3.3.4.1 Data

The method was tested on 2D cardiac MRI sequences obtained from Sunnybrook Cardiac Data [RLC09], which contains 45 4D short-axis cardiac cine MR scans, each containing 20 frames that cover the cardiac cycle. The image resolution was 256×256, with pixel spacing 1.25 mm, 10 slices, and slice thickness 8 mm. Segmentations of left ventricles were provided at end-diastole (ED) and end-systole (ES) frames. 45 4D scans were divided into training, validation, and testing sets, containing 30, 5, and 10 3D videos respectively. Fixed and moving image pairs were prepared by picking 2D slices from the same 4D scan, at the same slice position but at different time points in the cardiac cycle. Down-sampled in the temporal domain, 27,000 image pairs were used for training eventually.

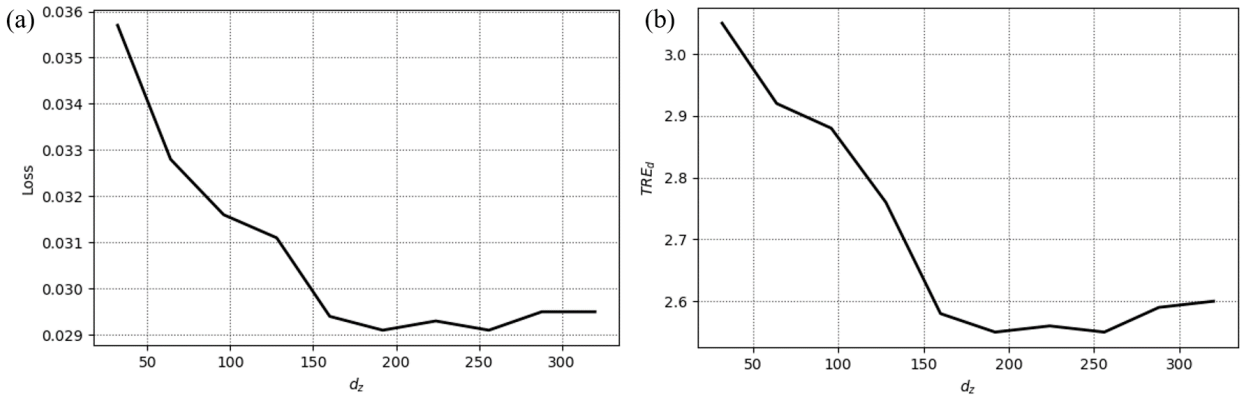

Figure 3.23: Results of the simulated CT registration with different latent variables lengths. (a) Final training loss. (b) Target registration error.

3.3.4.2 Model Training

The parameter σ in Eq. (3.10) was set to 0.002. The model was trained in mini-batches of 8 samples for 1500 epochs. The inferential back-propagation used 10 Langevin steps with the step size δ set to 0.003. The learning back-propagation used ADAM optimizer with the learning rate η set to 10^{-5}.

3.3.4.3 Evaluation

Using the provided segmentations of left ventricles, we computed the following metrics to evaluate our method against SimpleElastix with different trade-off parameters (weight for the bending energy λ set to 1, 5 and 25) and DIRNet (trained with and without bending energy penalty).

Dice coefficient between the propagated and fixed segmentation masks $M_{x'}$ and M_y was calculated:

$$Dice = \frac{2|M_{x'} \cap M_y|}{|M_{x'}| + |M_y|}. \tag{3.13}$$

Average surface distance (ASD) of the propagated and fixed segmentation contours was

calculated:

$$ASD = \frac{\sum_{a \in C_{x'}} dist(a, C_y) + \sum_{b \in C_y} dist(b, C_{x'})}{|C_{x'}| + |C_y|}, \tag{3.14}$$

where a and b are points on the propagated contour $C_{x'}$ and the fixed contour C_y, respectively.

Average foreground deformation magnitude (AFM), average background deformation magnitude (ABM) and their difference were calculated, with the foreground being the left ventricle and the background being outside of the mask:

$$AFM = \frac{1}{|M_y|} \sum_{(i,j) \in M_y} \|v_{ij}\|. \tag{3.15}$$

3.3.4.4 Result

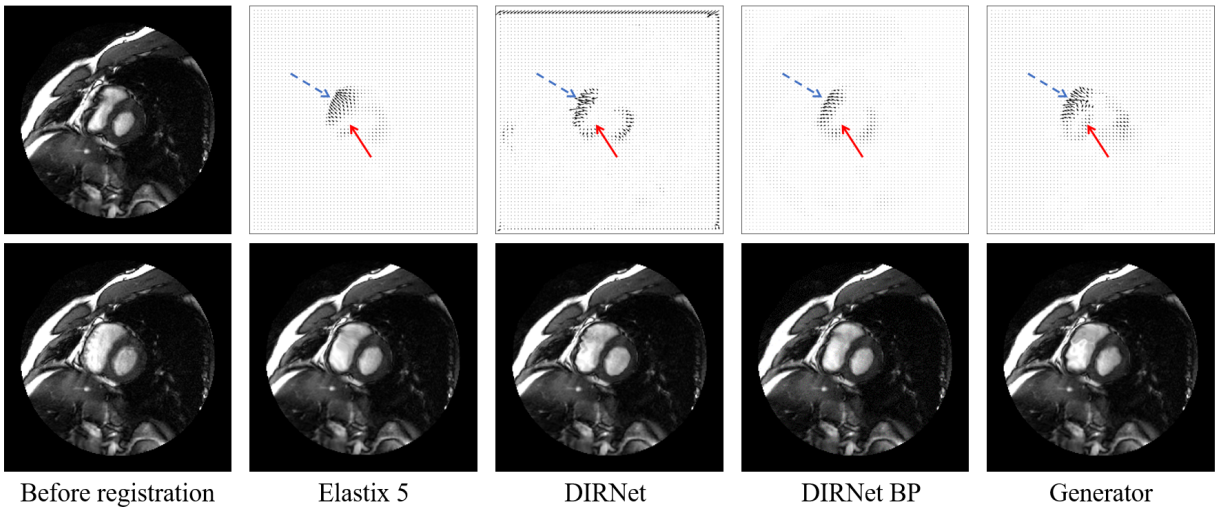

Figure 3.24: Example registration results on 2D cardiac MRI. Blue dashed arrows indicate the non-smoothness of the DVF. Red solid arrows indicate tissue movements in the myocardium region between the two chambers.

Fig. 3.24 shows a typical registration result. Upon close visual inspection, SimpleElastix generated a reasonable globally smoothed DVF, but imposing bending energy regularization at such level also compromised some local tissue movements as in the myocardium region between the two chambers. DIRNet as a single inference network offered a speed gain but

experienced non-smoothness of both the DVF and the warped image. While introducing bending energy penalty into DIRNet effectively remedied the non-smoothness, global vs local conflict persisted, as in the SimpleElastix model. Our method yielded DVFs that were better localized in the real motion area and thus physiologically more feasible.

Table 3.7: Results of the 2D cardiac MRI registration experiment. Results are provided as mean ± standard deviation (Z-value from Wilcoxon signed-rank test). Z-values that indicate statistical significance are underlined.

	Elastix1	Elastix5	Elastix25	DIRNet	DIRNet BP	Generator
Dice	0.90±0.03 (5.17)	0.92±0.03 (2.38)	0.92±0.03 (2.53)	0.90±0.04 (4.73)	0.91±0.03 (4.91)	**0.93±0.03**
ASD (mm)	2.13±1.29 (3.87)	1.65±0.98 (1.45)	**1.60±0.88** (1.60)	1.81±1.47 (3.63)	1.75±1.00 (2.68)	1.61±0.89
AFM (mm)	2.91±0.96	2.71±0.81	2.30±0.67	2.96±1.29	2.62±0.94	2.74±0.67
ABM (mm)	0.94±0.30	0.48±0.26	0.36±0.18	0.95±0.44	0.37±0.27	0.40±0.20
AFM−ABM (mm)	1.97±0.68 (3.77)	2.23±0.66 (1.47)	1.94±0.61 (3.89)	2.01±0.73 (2.28)	2.25±0.60 (1.59)	**2.34±0.57**
Time (s)	5.63			**0.005**		1.72

Table. 3.7 and Fig. 3.25 shows the quantitative results. Wilcoxon signed-rank tests were performed to examine the statistical significance (at the significance level of 0.05, Z_{crit}=1.96). In terms of dice coefficient, our generative model achieved the best result among all the methods tested, with statistical significance. In terms of ASD, our method outperformed DIRNet and was close to the best of the SimpleElastix results. The large gap between AFM and ABM in our method indicates that it focuses relatively large deformation in the foreground myocardial region and is more flexible in admitting spatial heterogeneity.

To further test the properties of the latent space, we examined the DVFs generated from interpolations of the latent variables z. Specifically, we first used the trained model for registration (a) between two frames near ES with small deformation (Fig. 3.26, 1st column), and (b) between ES and ED frames with large deformation (Fig. 3.26, 4th column). Then linear interpolations of the latent variables and the corresponding DVFs and warped images were generated, following the equation $z = (1 - \alpha)z_0 + \alpha z_1$. Results with $\alpha = 0.6, 0.8$ and corresponding nearest reference image frames are shown in the 2nd and 3rd column in

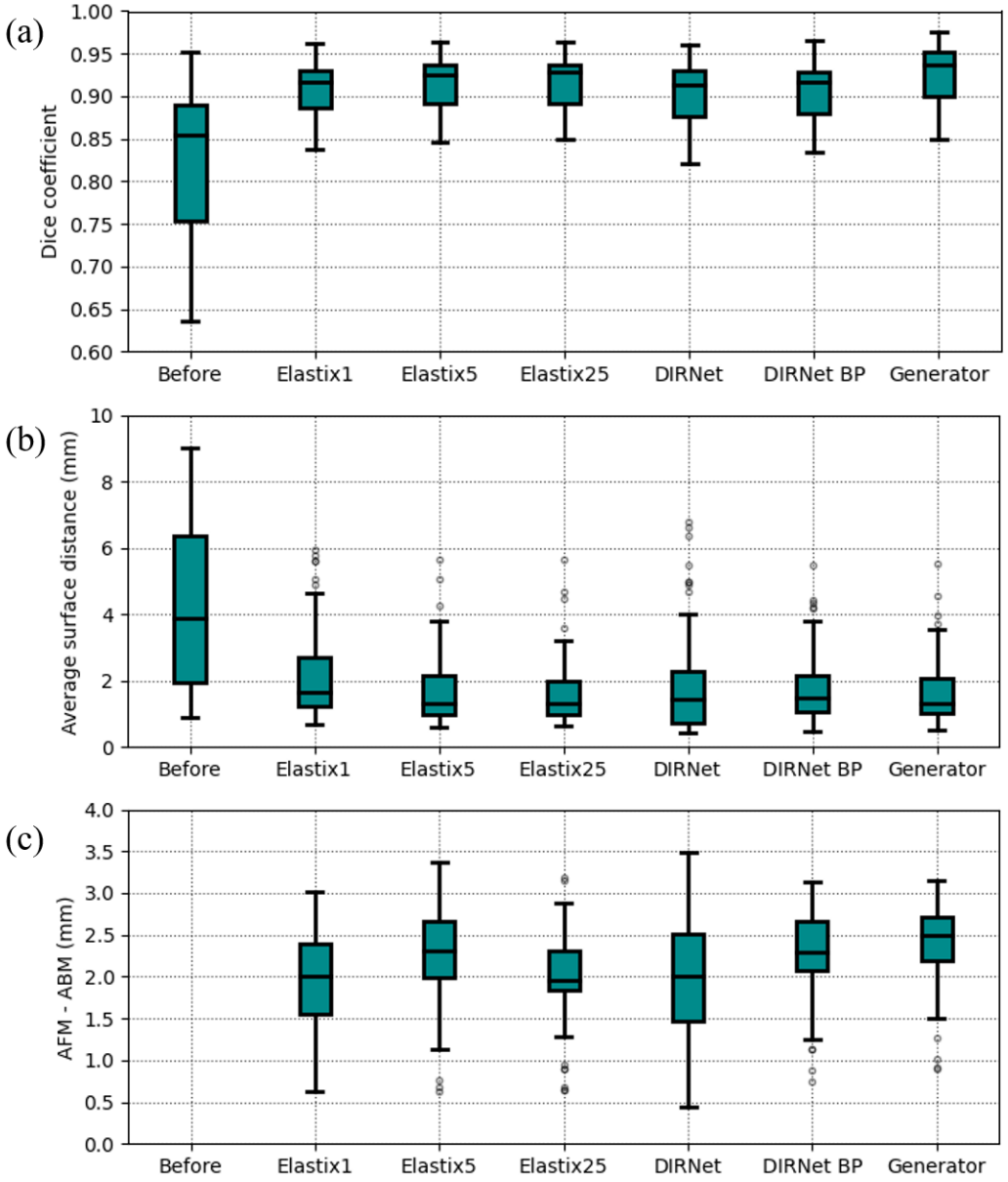

Figure 3.25: Boxplots showing the results of the 2D cardiac MRI registration evaluated with left ventricle segmentations. (a) Dice coefficient. (b) Average surface distance. (c) Difference between foreground and background deformation magnitude.

Fig. 3.26. It can be observed that the interpolations of z resulted in smooth transitions of DVFs and warped images. Although the linear interpolation of z did not precisely correspond

to the time step interpolation in the cardiac cycle, the result implied that the feasible DVF set was effectively characterized in the latent space with constraints.

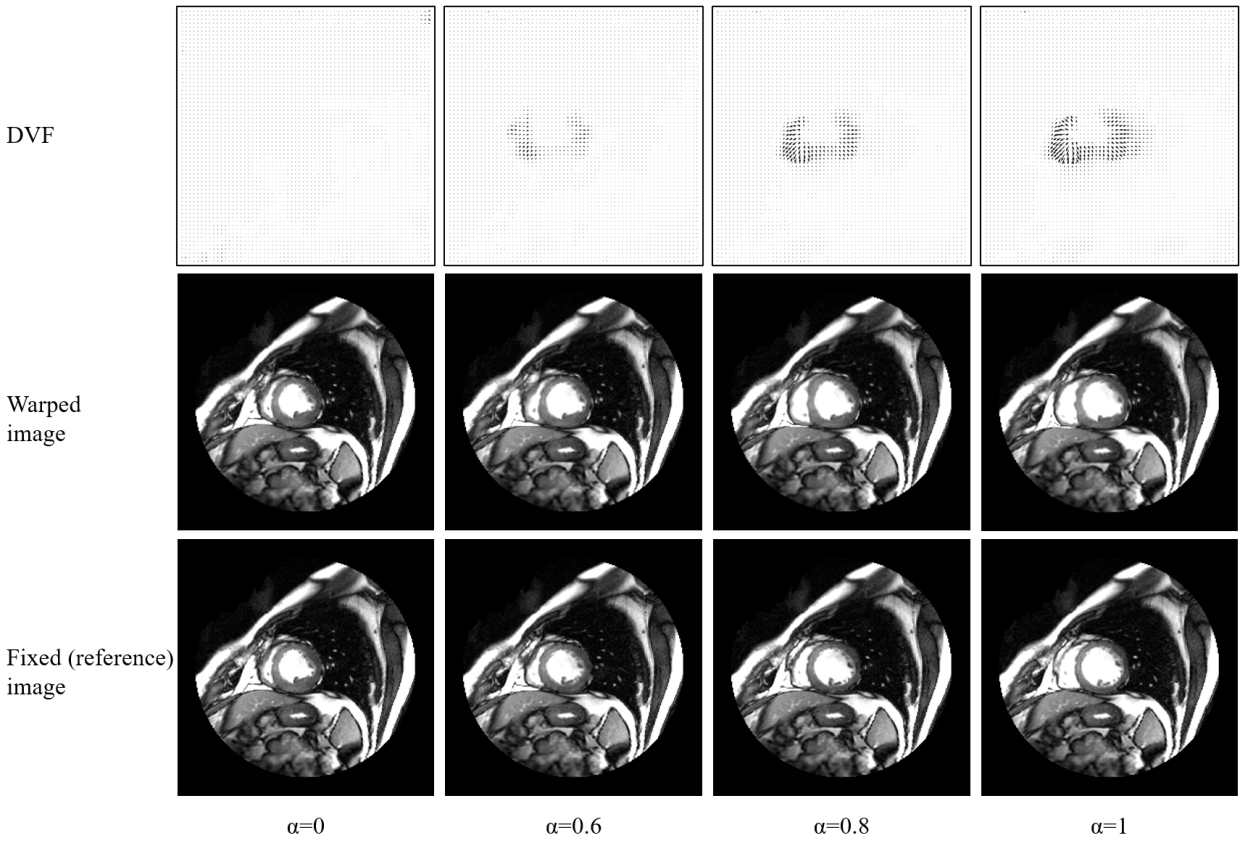

Figure 3.26: Results of the latent varibales interpolation experiment.

3.3.5 Experiments on 3D Cardiac MRI

3.3.5.1 Data

To account for realistic motions, the method was also tested on 3D cardiac MRI using the same dataset described in Sec. 3.3.4.1. 45 4D scans were divided into training, validation, and testing sets, containing 30, 5, and 10 3D videos respectively. Fixed and moving image pairs were from the same 4D scan, at different time points in the cardiac cycle. 11,400 3D image pairs were used for training.

3.3.5.2 Model Training

The parameter σ in Eq. (3.10) was set to 0.002. The model was trained with batch size of 1 for 1200 epochs. The inferential back-propagation used 10 Langevin steps with the step size δ set to 0.01. The learning back-propagation used ADAM optimizer with the learning rate η set to 10^{-4}.

3.3.5.3 Evaluation

Our method was evaluated using the left ventricle segmentations and was compared to 3D versions of B-spline registration in SimpleElastix, DIRNet, and VoxelMorph.

3.3.5.4 Result

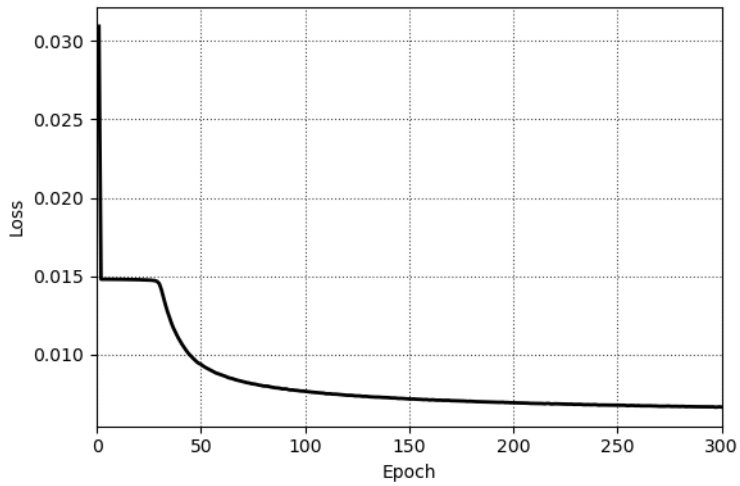

Figure 3.27: Learning curve showing the training loss for the 3D cardiac MRI registration. Only the first 300 epochs are shown.

Fig. 3.27 shows the learning curve. It can be observed that there was a platform at $loss = 0.015$ around the first 30 epochs. At this stage, DVF was equal to zero in every location (zero deformation). This platform corresponded to an internal rebalancing process

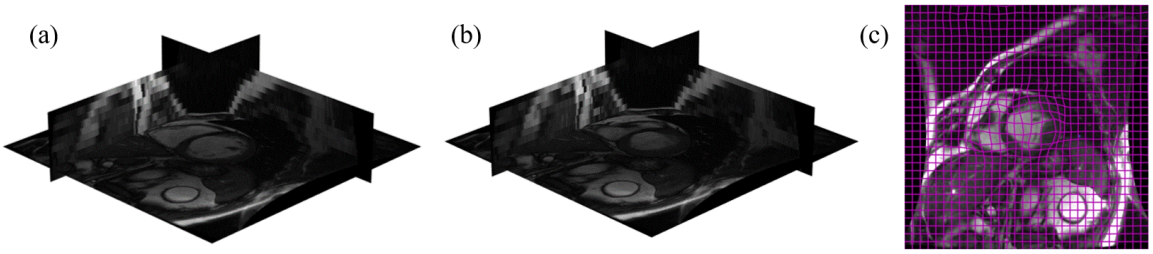

Figure 3.28: Example registration results on 3D cardiac MRI using our method. (a) Moving image. (b) Fixed image. (c) One slice from a 3D warped image. The DVF is illustrated with a mesh grid.

of z and θ. The zero-deformation platform existed in all the other experiments as well.

Fig. 3.28 shows a typical registration result. Table. 3.8 and Fig. 3.29 shows the quantitative results. Wilcoxon signed-rank tests were performed to examine the statistical significance (at the significance level of 0.05, $W_{crit} = 47$). The dice coefficient and ASD comparisons are generally consistent with the 2D experiment. Both VoxelMorph and our method outperformed DIRNet (with and without BP). Our method achieved a better registration result than VoxelMorph in Dice with statistical significance and in ASD without statistical significance.

Table 3.8: Results of the 3D cardiac MRI registration experiment. Results are provided as mean ± standard deviation (W-value from Wilcoxon signed-rank test). W-values that indicate statistical significance are underlined.

	Elastix30	Elastix65	Elastix100	DIRNet	DIRNet BP	VoxelMorph	Generator
Dice	0.90±0.02 (49)	0.92±0.02 (44)	0.91±0.02 (47)	0.91±0.02 (52)	0.91±0.02 (52)	0.92±0.03 (49)	**0.93±0.03**
ASD (mm)	2.01±1.04 (52)	1.33±0.65 (21)	**1.25±0.63** (15)	1.85±1.20 (50)	1.35±0.63 (20)	1.34±1.05 (32)	1.31±0.71
AFM (mm)	3.20±1.17	3.22±0.97	2.62±0.74	3.46±1.29	3.37±1.08	3.18±0.97	2.85±0.85
ABM (mm)	2.60±0.81	2.53±0.76	2.18±0.89	2.56±1.16	2.67±0.91	2.67±0.80	2.08±0.91
AFM−ABM (mm)	0.60±0.22 (51)	0.69±0.17 (46)	0.44±0.18 (53)	**0.90±0.41** (10)	0.70±0.20 (41)	0.51±0.23 (51)	0.77±0.34
Time (s)	24.70			**0.049**		0.065	12.78

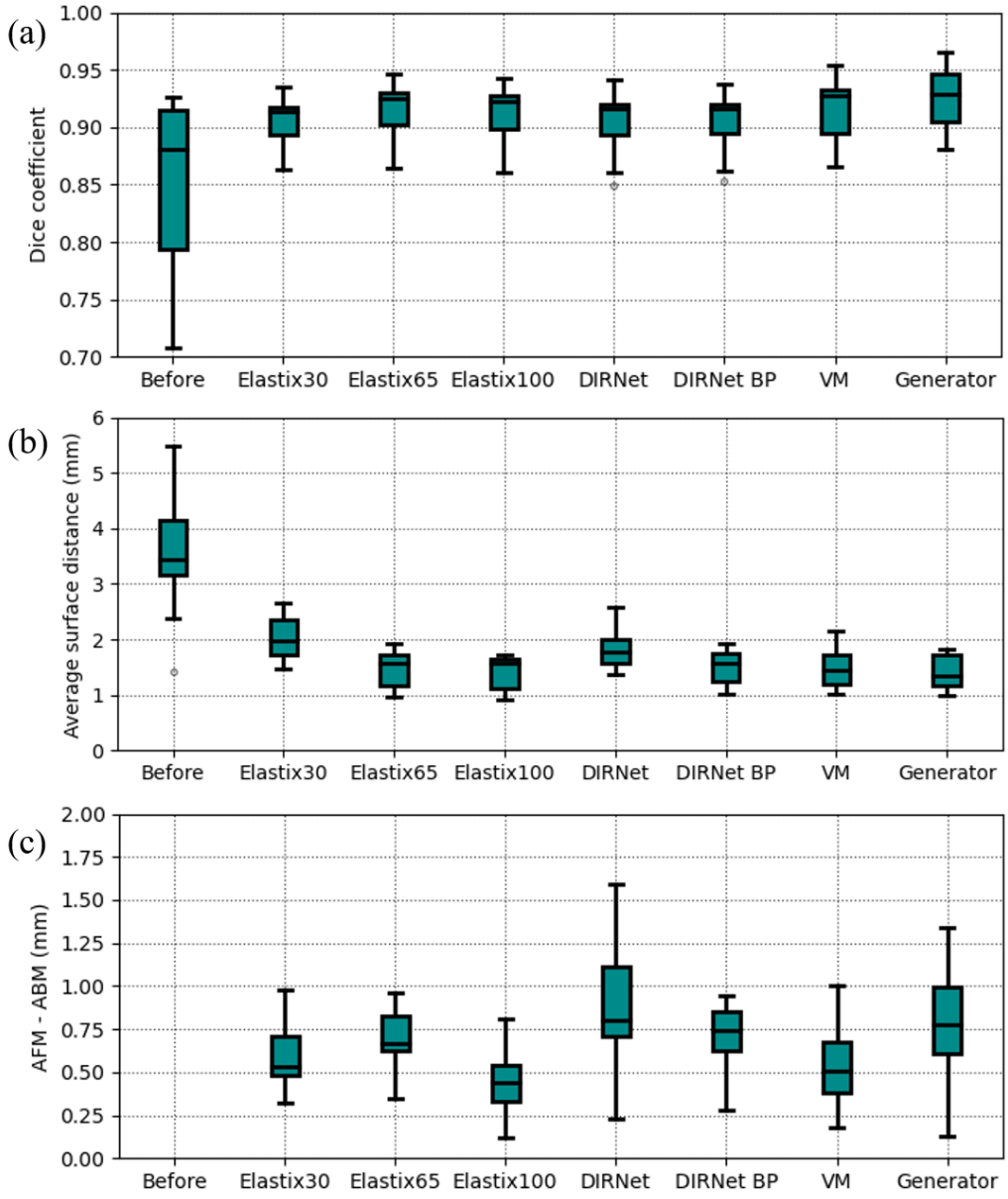

Figure 3.29: Boxplots showing the results of the 3D cardiac MRI registration evaluated with left ventricle segmentations. (a) Dice coefficient. (b) Average surface distance. (c) Difference between foreground and background deformation magnitude.

3.4 Discussion

In establishing performance benchmark with the classic BP regularized Elastic approach, we extensively traversed hyperparameter values and chose the one corresponding to the minimal

TRE from the test truth. Strictly speaking, this *Oracle* scheme of parameter setting provides a performance upper bound for the classic BP approach: even if the cost of running registration instances corresponding to different regularization weights can be accommodated, the absence of a ground-truth solution makes such choice infeasible in practice. The performance on the various learning-based methods reflects realistic behaviors.

In the proposed supervised approach, the DVF training samples were generated on high-quality images using a classic B-spline method in SimpleElastix with various values of registration weights. While the dataset contains sorting artifacts and the registration module is not ideal, we have observed that the use of various trade-off values and the CAE scheme for establishing feasibility descriptor is robust against moderate noise in the training DVF samples. The proposed CAE-based feasibility prior model is spatially variant and driven by the training DVFs, which enables it to better accommodate site-specific heterogeneous tissue properties and motion patterns than other "prescription-type" priors. We have also demonstrated the solution stability of the overall unsupervised registration network upon incorporating the CAE prior, alluding to the feasibility for direct use without further tuning. Furthermore, the proposed prior can be combined or incorporated into most registration networks, with any choice of image dissimilarity loss and network architecture (e.g. networks with inverse-consistency [Zha18], diffeomorphism [DBG19]), or discriminating adversarial block based on biomechanical modeling [ZHW19]. In the experiments on lung images, scans from patients with thoracic malignancies are used for both the feasibility descriptor and registration network training. In the cardiac experiment, the feasibility descriptor was trained on patients with cardiovascular deceases, while the registration was tested on both healthy and pathological cases. The current study diversified the DVF training set by varying the hyper-parameters in classic registration. To build a reliable prior and ensure the method's stability in clinical applications, the DVF training set should be large and diverse enough to provide proper coverage of representative patient and possible abnormalities.

While the supervised approach uses a DVF set acquired with another registration engine

to train the feasibility prior, it differs from supervised learning-based registration in that the training does not have to be consistent with the inference setup of image input, nor does it involve any image synthesis. In fact, the decoupled prior generation and registration network training is a major distinctive feature of our development, so that the feasibility prior can take advantage of high-quality CT for the lung cases studied here. And for that matter, the feasibility descriptor can be developed with any other image modality that could offer benefit in any of the resolution, signal strength, consistency, or contrast aspects. This decoupling not only allows the development of prior to be dependent on another data set and modality, but also permits the prior to be based on DVFs generated with a much wider range of options, in contrast to the cooperative prior [BEK19]. Offline CT registration can afford to be performed with more sophisticated methods at a higher time cost, even with manual interaction or ROI/contour/landmark guidance when needed, to ensure the quality of feasibility prior. Therefore, the proposed approach in developing and using the feasibility prior enjoys the benefit from both data quality and registration quality aspects. Compared with the adversarial network discriminating between predicted DVFs and biomechanical model-based motions [HGG18], the proposed network is more stable and simpler in the training setup. Once properly trained, the feasibility prior can be used as a PnP module for the re-training of the unsupervised DVF estimation network as it migrates to new imaging platforms, transferring the advantages of prior-generating quality and fine-tuned methods into the inference imaging modality. This flexibility further allows the end-users to choose either utilizing an existing feasibility descriptor with applicable site or customizing the prior training with their best available dataset and registration resource.

In the unsupervised approach, the heterogeneous feasibility prior is achieved by constraints on the latent variables, rather than the DVF itself. This allows the model to reallocate the "budget" of constraint through the unsupervised training process, with the generator network serving as a mapping between the feasible z set and feasible DVF set. The learned implicit parametrization is flexible in admitting spatial heterogeneity. Compared to the su-

pervised method, its training does not depends on any DVF generation approach and is therefore unbiased when high-quality images are available for training.

The latent variables length sensitivity test has demonstrated that the model performance is stable and robust to small-to-medium variations in d_z, and that the training does not require subjective selection among multiple solutions as in conventional methods. However, it should be noted that using too large d_z may require a longer training process and also increase the chance of local minimum due to over-fitting.

The current study focus on inter-phase DIR of a 4D image, where DVFs reflect tissue motion and the feasibility priors describe a manifold of possible motions. The proposed methods generalize to other settings. For example, when performing registration on 3D images across the timeline of radiation therapy, a feasibility prior can be trained in the same setting to model longitudinal change and characteristics of anatomy during the course of of radiation therapy.

3.5 Conclusion

We have developed two learning-based methods for DIR. In the supervised method, a PnP feasible motion prior is developed from high-quality images, and then incorporated as a regularizer to train an unsupervised DIR network. In the unsupervised method, novel deformation parametrization in the form of a generator network is developed to learn implicit feasibility conditions on DVF from paired images using the alternating back-propagation algorithm. During registration, the latent variables are optimized, eliminating the need for regularization and tuning. The methods managed to model DVF feasibility conditions effectively, which is one of the major challenges in DIR due to the heterogeneous tissue and motion properties. Both methods have yielded promising results with high accuracy and efficiency. The two methods can complement each other, and their application depends on whether a reliable feasible DVF sample generation method is available in the specific task.

CHAPTER 4

Individualized Test-time Adaptation in Deformable Image Registration

4.1 Introduction

DL approach faces a particular challenge in generalization. In the context of DIR, the generalization challenge manifests in various aspects. It may risk generating inferior results when the testing data deviates from the training probability distribution, in either image characteristics or motion patterns or both. Here, we refer to the input-space image characteristics variations as **domain shift**, and the output-space DVF characteristics variations as **generalization gap**.

Currently, most existing studies circumvent the generalization challenge by using image datasets acquired on the same scanner with the same imaging protocol and then dividing them into training and testing subsets to ensure the consistency of image characteristics. Typically, further care is also taken in the division to intentionally match the training and testing cohort distribution with respect to demographic, disease grading, or histopathology etc. However, in real-world scenarios, the testing samples may be acquired with scanners parameters, imaging protocols, contrast agent, and preprocessing standards that are either (1) significantly different than the data cohort used for training and establishing the deep network model or (2) different between different cases at testing, or (3) both. In addition, individuals may have large differences in physiological and pathological presentations, particularly in applications where we aim to diagnose or treat ab-

normalities. It is typically challenging to cover all variations within the training dataset. Although data augmentation techniques can be used to generate training samples in large quantity [EP18, ZWY20, SPC18, HBG21], it is difficult to achieve a good balance between (1) generating feasible and realistic images and DVFs and avoiding aggressive augmentation to introduce artificial samples and (2) generating a rich set of samples to cover sufficient variability. The application of DL methods on new or less common modalities/protocols is also challenging without a sufficiently large training cohort [WL16].

The robustness of DL methods is an important factor that challenges their practical use in clinical routines. Most DL methods are trained to optimize the average performance on a certain cohort. However, ensuring that the network has good performance on each single test sample, in other words, increasing the performance lower bound, has been a widely unaddressed problem. In clinical applications, the failure of DL methods on individual cases may cause severe consequences, despite the average performance. A training scheme that aims at improving individualized performance or worst-case guarantee is in demand [VQL20].

In this study, we propose a domain adaptation method in DL-DIR to address the potential domain mismatch between training and testing images, and improve the accuracy and robustness of registration. Specifically, we propose a test-time refinement training scheme to adapt a DIR network to individual test image characteristics and motion details.

4.2 Related Works

4.2.1 Generalization Gap – Output-space DVF Variations

Generalization gap is usually defined as the difference between a model's performance on training data and its performance on unseen data. Understanding and reducing the generalization gap have a great practical significance and have been studied extensively.

In classification, detection tasks, generalization robustness is mainly challenged by label

class imbalance. Specifically, neural networks tend not to effectively model underrepresented classes, which can affect its fairness towards testing samples that are underrepresented in the training set. In these tasks, the model's performance is usually examined on a cohort basis, by calculating average performance metrics across all classes. Many training schemes have been proposed to address the class imbalance problem. The training dataset can be rebalanced by discounting samples in the majority classes, replicating samples in the minority classes, or simulating synthetic samples for the minority classes for augmentation [HG09, HLL16, CBH02]. Other studies modify the loss function and re-weight training samples with the inverse label frequency [Jap00].

The class imbalance problem generalizes to the out-of-distribution DVFs in DIR. The characteristics of the DVF can differ significantly across different individuals, and it is difficult for the training cohort to properly cover all sexes, ages, and possible abnormalities. Furthermore, even if a large and diverse training cohort can be obtained, individual-level anatomical and functional variations still challenge the registration accuracy in relatively underrepresented cases.

DIR performance should be quantified not only on a cohort basis, but also on an individual basis, by calculating the target registration error. In addition to the average performance on the testing cohort, it is arguably even more important to improve the DIR performance lower bound with respect to individual-specific cases, especially in applications where we aim to diagnose or treat abnormalities. While the cohort-level robustness can be addressed using similar rebalancing approaches as in the classification tasks, the individual-level robustness in DIR has rarely been studied.

4.2.2 Domain Shift – Input-space Image Variations

Domain shift is a universal challenge that may comprise performance in almost all deep learning image processing tasks. In the context of DL-DIR, domain shift refers to the image appearance and characteristic differences between the domain of training samples and the

domain of testing samples – variations in image intensity statistics and contrasts between different tissue types. Such difference can occur as a result of applying a trained network to images in a different modality, acquired with a different imaging protocol or parameters, or from a different scanning platform with software or hardware [MNJ20, PKH20]. Since neural networks make predictions based on the learned input-output mapping, such variations in image characteristics at test time can significantly degrade the network performance.

Different training schemes have been proposed to address the domain shift problem. The most direct approach to encourage domain invariance is to incorporate the possible variations in the training set. When training samples for new or less common modalities/protocols are difficult to obtain, simulation can be used for data augmentation [ZWY20, SPC18, JHG19, HBG21]. Another approach is to utilize domain-invariant features that can be obtained from a separate pre-training step using autoencoders that reconstruct images from various domains [GKZ15]. It is also possible to enforce feature similarity across all domains during the training process using a regularization loss to enforce such consistency [MPA17].

A common disadvantage of these methods is that they usually require samples from the two domains to be both present in the training process. However, in real-world scenarios, samples in the transferred domain can be difficult to obtain or even unknown when training the network. Therefore, a test-time adaptation process that is separate from the training is preferred.

4.2.3 Test-time Adaptation

Test-time adaptation can potentially address both the generalization gap in output space and the domain shift in input space at the same time. In addition, the adaptation can be further individualized to improve the targeted performance on each testing image pair or sequence. While there is relatively limited amount of investigations on the generalization and adaptation issue on DIR, investigation efforts have been made on segmentation and reconstruction in recent years. A pre-trained network can be adapted to each test case in a refinement

training process formulated in an unsupervised setting. Approaches with the general test-time adaptation rationale has been explored recently in segmentation, reconstruction, and classification works, generating promising results.

Karani et al. [KEC21] designed a segmentation network as a concatenation of two sub-networks: a shallow image normalization network, followed by a deep network that segments the normalized image. The two sub-networks are trained jointly in a supervised setting. At test-time, the normalization network is adapted to each testing image, guided by the auto-encoding discrepancy of the segmentation provided by a pre-trained denoising autoencoder (DAE) that predicts clean labels from artificially corrupted segmentations. Zhang et al. [ZLZ20] proposed to fine-tune a supervised pre-trained reconstruction network for each testing case by minimizing an unsupervised fidelity loss function defined according to a forward physical model of the imaging system and data noise property. Wu et al. [WKL21] proposed a reconstruction algorithm where a Noise2Noise network is employed as image prior. The network is fine-tuned along with each image during the iterative reconstruction with an alternating optimization strategy. Sun et al. [SWL20] proposed to adapt part of an image classification network according to a self-supervised loss defined on the given test image. Specifically, they used the task of predicting a four-way image rotation angle as an auxiliary task and utilized its self-supervised loss to adapt the network. All these methods rely on an unsupervised loss to guide the adaptation process, by modeling the feasibility of the prediction or from an auxiliary task.

DIR can be formulated as an unsupervised task, making it natural to utilize the self-supervised image dissimilarity loss for individualized network fine-tuning. However, domain adapting in DIR has rarely been studied. Zhu et al. [ZHX21] proposed to further optimize a trained DIR network based on both training set and testing image pairs to reduce the generalization gap with multiple experiments. However, with training and testing images drawn from the same datasets, the domain shift aspect is not sufficiently addressed or validated.

4.3 Method

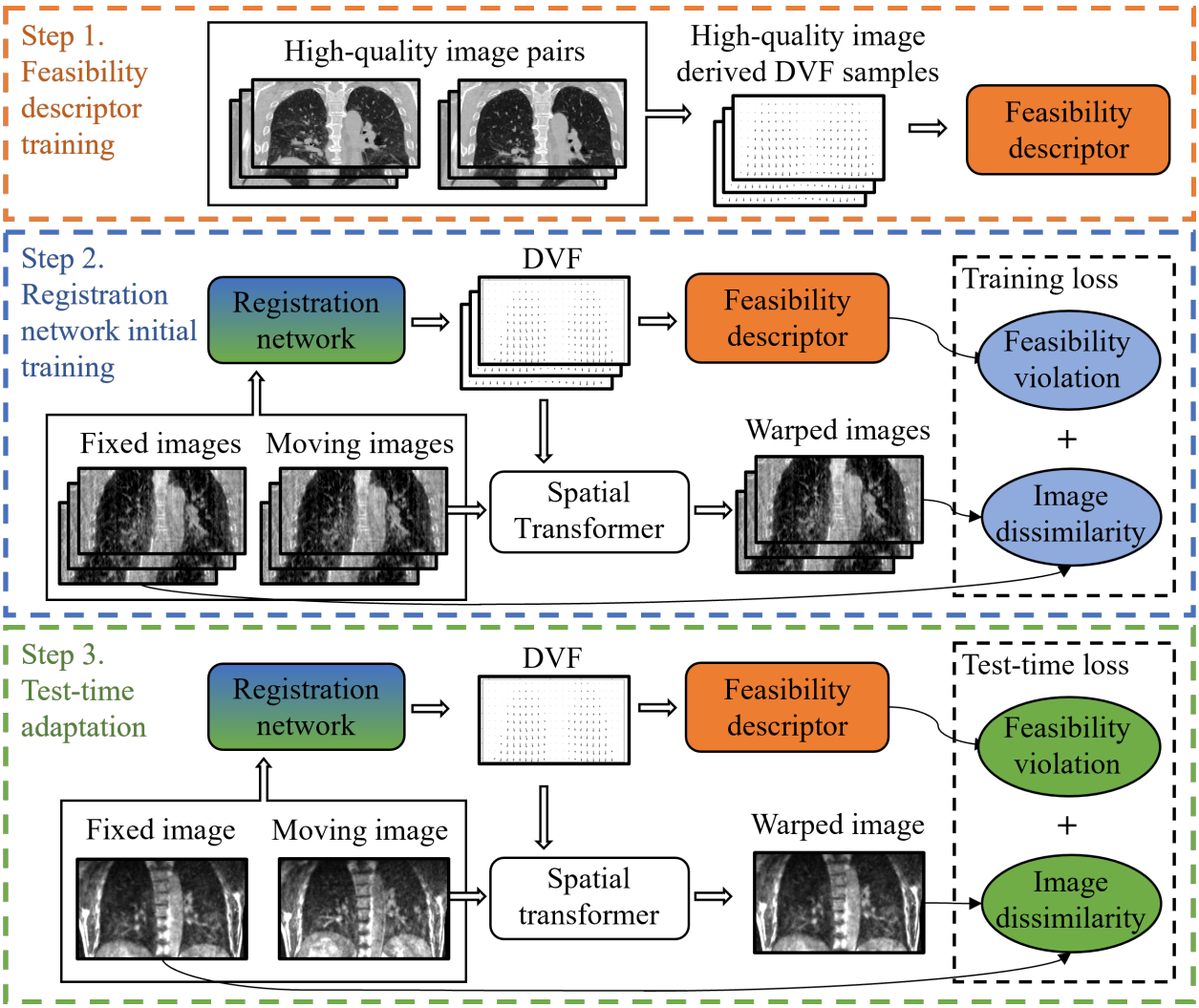

Figure 4.1: Overview of the proposed method. The feasibility descriptor is pre-trained and used as a regularizer during the subsequent registration network training. The registration network is initially trained on a set of image pairs, and then refined and adapted to individual image pair in another domain at test time.

4.3.1 Baseline Registration Network

The current development extends from the unsupervised DIR network with spatially variant, modality-agnostic DVF feasibility prior [SCM21], as shown in Fig. 4.1 (step 1 and 2). The feasibility descriptor and the registration network are developed sequentially. In the form of a CAE, the feasibility descriptor encodes Jacobian properties of training DVFs with an L2 objective:

$$L_{CAE} = \frac{1}{N}\sum_{i=1}^{N} \| \nabla(v_i) - \mathrm{CAE}(\nabla(v_i))\|^2_{\mathrm{Frob}}, \tag{4.1}$$

where v_i is a DVF training sample, N is the number of samples, and ∇ is the Jacobian operator. Once the CAE is properly trained, the deviation of a candidate DVF from feasibility is described by the auto-encoding discrepancy. After training, the feasibility descriptor is fixed and plugged into the registration network as a regularizer.

The registration network has a U-net structure and was trained to minimize a loss consisting of a matching cost L_D, where in this specific setting, we specialized to the form of NCC, and a feasibility violation penalty L_V. The loss function can be written as:

$$L_{train} = L_D + \mu L_V = \frac{1}{P}\sum_{i=1}^{P} \left\{ -\mathrm{NCC}(F_i, M_i \circ v_i) + \mu \| \nabla(v_i) - \mathrm{CAE}(\nabla(v_i))\|^2_{\mathrm{Frob}} \right\},$$

where P is the number of samples, (F_i, M_i) is a fixed and moving image pair, $v_i = \mathrm{UNet}(F_i, M_i; \theta)$ is the DVF estimated by the registration network parameterized by θ, and μ is a balancing parameter.

4.3.2 Test-time Adaptation

In contrast to most existing DL-DIR methods where the DVF is inferred directly by applying the trained network to the testing inputs [BZS19, VBV19], we propose an individualized test-time adaptation scheme to mitigate the potential risk of sub-optimal or biased DVF estimate when the test sample deviate from training distribution, in terms of image characteristics and motion patterns.

Specifically, to inform the DIR network of test sample characteristics, we introduce a test-time refinement process to further evolve the network with respect to an individualized loss by taking advantage of the unsupervised framework and treating the typical DL-stage as prior transfer. At test time, the image pair is used to train the registration network until convergence with respect to a new test-time loss. The feasibility descriptor remains fixed during the adaptation process to guarantee stability.

$$L_{test} = -\mathrm{NCC}(F_t, M_t \circ v_t) + \mu \| \nabla(v_t) - \mathrm{CAE}(\nabla(v_t)) \|^2_{\mathrm{Frob}},$$

where (F_i, M_i) is the test image pair, $v_t = \mathrm{UNet}(F_t, M_t; \theta)$, and the network parameter θ is initialized from the cohort training stage and further optimized.

4.4 Experiments and Results

Table 4.1: Experimental setup.

	Site	Prior generation	Initial training	Test sample
Cross-protocol	Lung	CT	CBCT	Simulated CBCT
Cross-platform	Heart	CTA	0.35T MRI	1.5T MRI
Cross-modality	Lung	CT	CBCT	MRI

The method was tested in three experiments that correspond to cross-protocol, cross-platform, and cross-modality scenarios, respectively, as indicated in Table.4.1. Its performance was compared against a classic B-spline method in SimpleElastix [MBS16] and a benchmark registration network inference without test-time adaption. The classic B-spline registration was driven by weighted sum of NCC and bending energy penalty (BP), with the weight for BP manually tuned for each case. ADAM optimizer with learning rate 10^{-4} was used for network training throughout all experiments.

To derive realistic DVF samples for feasibility descriptor training, the classic B-spline registration was performed on high-quality images in each experiment. In order to accom-

modate spatially variant regularization, various values of regularization weights λ was used, so that the learned manifold could address different local trade-offs.

4.4.1 Cross-protocol Adaptation on Lung CBCT

4.4.1.1 Data and Network Training

Feasibility descriptor training: Classic B-spline registration on a set of 10 CT images from the DIR-Lab dataset [CCG09, CCM09]. The slice thickness was 2.5 mm and in-plane spacing was 0.97 to 1.16 mm. All images were resampled with slice thickness 2.34 mm and in-plane pixel spacing 1.16 mm, and then cropped to a $256 \times 256 \times 64$ window that covered the lungs. Image intensities were clamped between -1000 and 500 HU and scaled between 0 and 1.

The BP weights λ was set between 0.01 to 2. For each of the 10 scans, 15 moving and fixed image pairs were selected. Then, they were augmented by 5 registrations performed with different λ. As a result, 750 DVFs were generated as the training set. The CAE was trained for 200 epochs with batch size 1.

Registration network training: The 4D-CBCT data was from the 4D-Lung collection in the Cancer Imaging Archive (TCIA) [HWS17]. They were acquired during chemoradiotherapy of 20 locally advanced, non-small cell lung cancer patients. Each scan has 10 breathing phases. The reconstructed slice thickness was 3 mm and in-plane spacing was 0.98 to 1.17 mm. The images were pre-processed to the same size and pixel spacing as the CTs. 25 scans were used for training. For each scan, 15 moving and fixed image pairs were selected. The network was trained for 150 epochs with batch size 1. The balancing parameter μ in Eq. (4.2) was tuned based on the training performance, and was set to 10^{-6} subsequently.

Test-time adaptation: The cross-protocol adaption was tested on a set of 4D-CBCT scans simulated from CT scans in the SPARE dataset [SGL19]. Each phase in the CBCT was simulated independently, using the corresponding 3D CT data. The geometry and

the number of projections were set according to a typical clinical setup and is different from the TCIA data. FDK reconstruction algorithm was then applied using the TIGRE toolbox [FDK84, BDH16]. Eventually, nine scans from the dataset were used. At test time, the network was trained with the individual CBCT scans for 200 epochs.

4.4.1.2 Evaluation

We applied an automatic landmark pair detection algorithm [FWT19] to the original CTs in the SPARE dataset to take advantage of its higher image quality and structural details. The locations of the landmarks were then mapped to the simulated CBCTs. Each scan had 100 landmark pairs in the EI and EE phases. TRE defined by the Euclidean distance between the transformed and the fixed landmarks was calculated as to measure registration performance.

An accurate DIR should be able to identify the motion trajectory of each pixel, and integration along such trajectory can enhance image quality. A simple motion-compensated image enhancement test was performed by collapsing all phases of the 4D-CBCT according to the estimated DVFs to an arbitrary reference phase and taking the average. The enhancement is quantified with RMSE and structural similarity index measure (SSIM) to the ground-truth CT. Paired t-tests were used to examine statistical significance.

4.4.1.3 Result

To illustrate the behavior of the test-time adaptation, the loss curves showing the transition between the train and adaptation stages are shown in Fig. 4.2. A large generalization gap between the training cohort and the test sample can be observed in the training stage. The individualized adaptation successfully fitted the model to the test sample and reduced the test-time loss.

As shown in Fig. 4.3, all three methods generated relatively smooth DVFs. However,

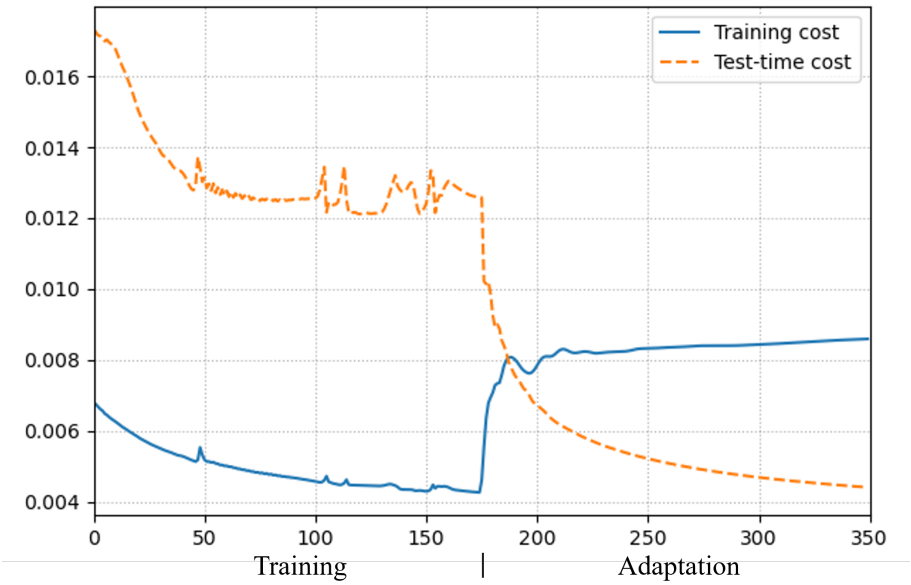

Figure 4.2: Loss curves for the cross-protocol adaptation on lung CBCT.

classic B-spline and the network without adaptation were affected by the noise and artifacts in the lung regions and underestimated the motion in homogeneous regions. The test-time adaptation preserved the overall motion and resulted in a smoother DVF.

As shown in Fig. 4.4, in the CBCT enhancement experiment, the streak artifacts were less pronounced in the fusion results from the network with and without adaptation, as indicated by the red boxes. Because of the higher spatial accuracy, the test-time adaptation achieved higher visual resolution and managed to reconstruct sharper detail structures, as indicated by the red ovals.

Table. 4.2 shows the quantitative results. Our method achieved the best TRE, RMSE, and SSIM, with statistically significant ($p < 0.01$) improvement over the other two methods. The average registration time was 52, 0.04, and 21 s for classic B-spline, network without adaption, and network with adaption, respectively.

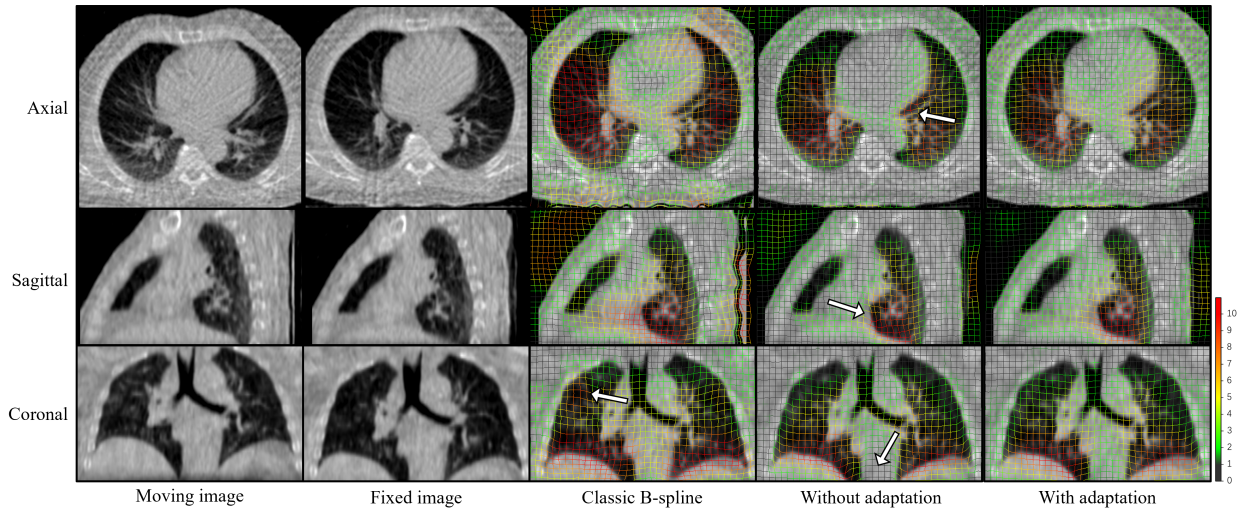

Figure 4.3: Example registration results on simulated lung CBCT.

Table 4.2: Quantitative results on simulated lung CBCT. Results are provided as mean ± standard deviation (p-value from paired t-tests).

	TRE (mm)	RMSE (HU)	SSIM
Before	7.53±4.15	150.2±11.85	0.982±0.004
Classic B-spline	2.55±2.45 (10^{-3})	115.3±7.83 (10^{-7})	0.991±0.003 (10^{-3})
Without adaptation	2.13±1.84 (0.005)	107.5±8.51 (10^{-3})	0.993±0.002 (0.002)
Our method	**2.11±1.61**	**102.1±7.96**	**0.994±0.002**

4.4.2 Cross-platform Adaptation on Cardiac MRI

4.4.2.1 Data and Network Training

Feasibility descriptor training: The CTA dataset for DVF sample generation consists of 10 4D scans, each containing the ED and ES frames of a cardiac cycle. The scans used contrast according to typical clinical system, on patients who are suspected to have cardiovascular problems. The image size were 512×512, with number of slices ranging from 240 to 564, in-plane resolution ranging from 0.31 to 0.45 mm, and slice thickness ranging from 0.30 to 0.50 mm. The images were resampled with voxel spacing 0.5 mm and cropped with

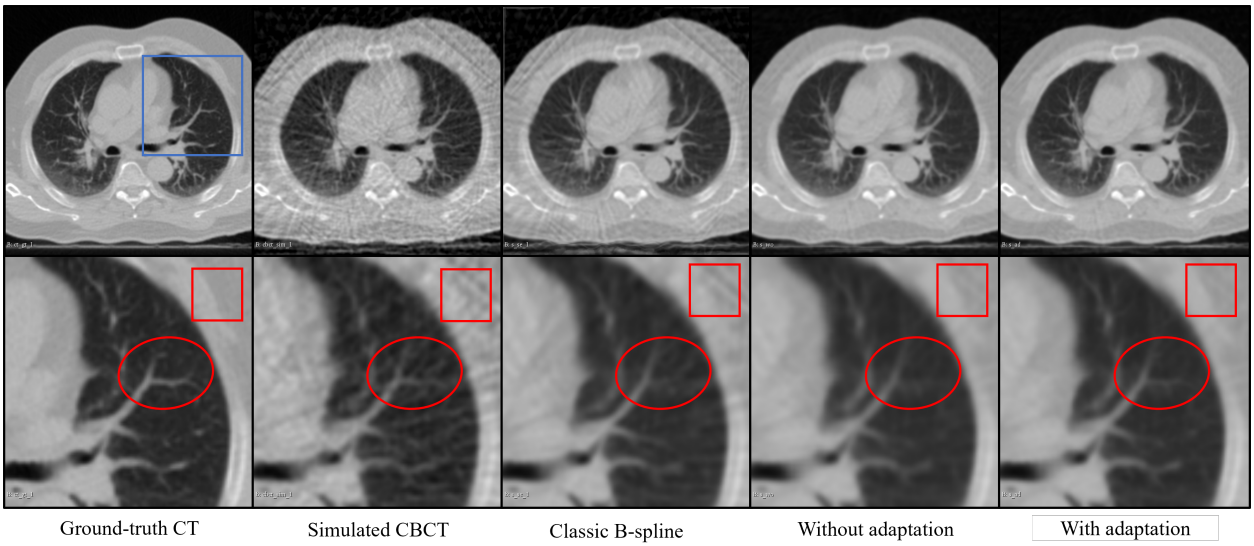

Figure 4.4: Example motion-compensated image enhancement results on simulated CBCT. The streak artifacts are better alleviated in the network results. Sharper detailed structures are reconstructed after adaptation.

a $320 \times 320 \times 224$ window that covered the entire heart.

The BP weights λ was set between 0.01 to 2. All 10 scans were registered 10 times, using different BP regularization weights, giving rise to 100 DVFs as the training set. The network was trained with batch size of 1 and 1200 epochs.

Registration network training: The registration network was trained on nine 0.35T MRI scans acquired on a ViewRay MRIdian system from healthy volunteers, with a breath-hold and EKG gating protocol, each consisting of ED and ES frames from balanced steady-state free precession (bSSFP) sequence. The slice thickness ranged from 6.00 to 8.00 mm with an average slice number of 25. The in-plane pixel spacing ranged from 1.25 to 2.18 mm. Each image was resampled with voxel spacing of 1 mm and cropped to $160 \times 160 \times 112$. The image intensity was cropped between 0 and 1000 and then normalized to [0,1].

The network was trained for 2000 epochs with a batch size of 1. The balancing parameter μ was tuned based on the training performance, and was set to 10^{-7} subsequently.

Test-time adaptation: The cross-platform adaption was tested on 15 1.5T MRIs in the cMAC public dataset [TDM13]. The scans were in short-axis view and each had 30 frames. The image size was typically $256 \times 256 \times 14$, with voxel size $1.25 \times 1.25 \times 8$ mm. The image was resampled to horizontal long-axis-view grid, with 1 mm voxel spacing. Then, they were cropped and normalized to the same size as the 0.35T images. At test time, the network was trained with the individual scans for 2000 epochs.

4.4.2.2 Evaluation

In each 1.5T MRI scan in the cMAC dataset, 12 landmarks on left ventricle were located using the corresponding tagged MRI: one landmark per wall (anterior, lateral, posterior, septal) per ventricular level (basal, midventricular, apical). The landmarks were manually tracked by two observers. The median of the inter-observer variability was 0.84mm. Euclidean distance between the fixed and the transformed landmarks was TRE to measure registration performance. Paired t-test was performed to examine the performance improvement.

4.4.2.3 Result

The loss curves showing the transition between the train and adaptation stages are shown in Fig. 4.5.

As shown in Fig. 4.6, the classic method generated an overly smoothed DVF, with unsatisfactory shape and intensity matching. The test-time adaptation successfully refined the local motion field and made it physically more feasible.

Fig. 4.7 shows an example where our method failed to estimate the left ventricle motion correctly. In this instance, the image intensity gradient was too strong for the motion prior to overcome, and the adaptation process did not reduce the DVF feasibility loss significantly.

As shown in Table. 4.3, our method significantly reduced the TREs compared to the network without adaption, and achieved comparable result to the carefully tuned classic

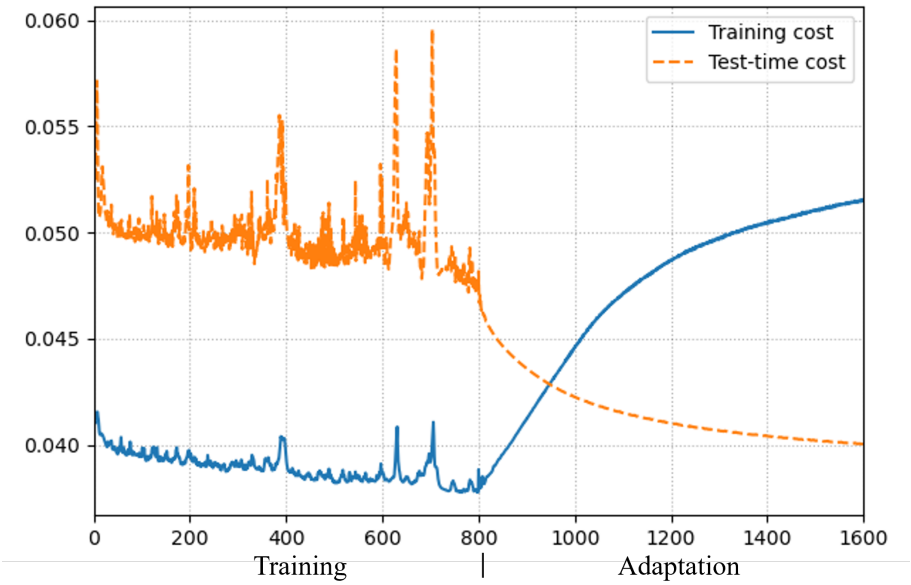

Figure 4.5: Loss curves for the cross-platform adaptation on cardiac MRI.

method.

4.4.3 Cross-modality Adaption on Lung MRI

4.4.3.1 Data and Network Training

Feasibility descriptor and registration network: In this experiment, the pre-trained networks were kept the same as 4.4.1.1. I.e., the feasibility descriptor was trained on CTs and the registration network was trained on CBCTs.

Test-time adaptation: The cross-modality adaptation was performed on a lung 4D MRI scan for 300 epochs. The scan had 8 phases. The image size was $264 \times 384 \times 112$, with voxel size $1.22 \times 1.22 \times 1.80$ mm. It was pre-processed to the same size and pixel spacing as the CTs in Sec.4.4.1.1.

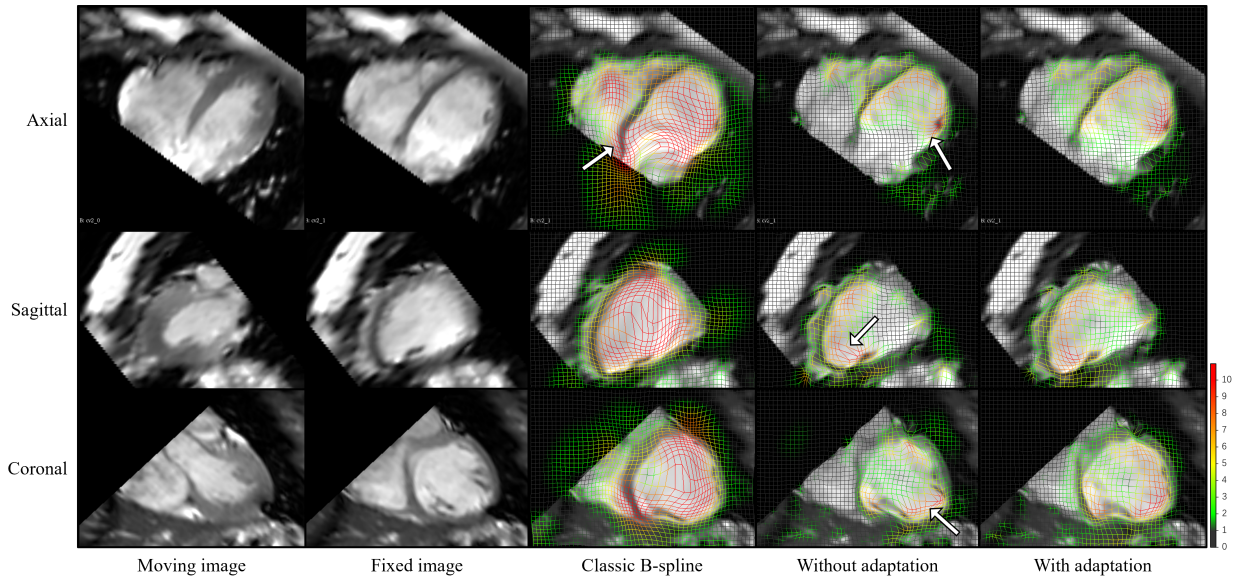

Figure 4.6: Example registration results on cardiac MRI.

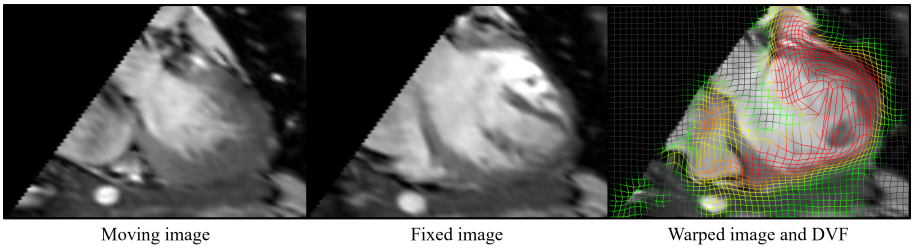

Figure 4.7: An example registration result on cardiac MRI where our method failed to correct the DVF.

4.4.3.2 Evaluation

We annotated 20 anatomical landmark pairs at EI and EE phases. The Euclidean distance between the transformed and the fixed landmarks was calculated as TRE to measure registration performance. Paired t-tests were used to examine statistical significance.

Table 4.3: Target registration errors based on landmarks on cardiac MRI. Results are provided as mean ± standard deviation in millimeter. Values inside the parentheses indicate p-value results from paired t-tests when comparing to our method. p-values that indicate statistical significance (< 0.01) are underlined.

	Basal	Midventricular	Apical	All
Before	8.73±2.20 ($\underline{10^{-25}}$)	7.82±2.90 ($\underline{10^{-15}}$)	7.30±3.24 ($\underline{10^{-20}}$)	7.95±2.88 ($\underline{10^{-52}}$)
Classic B-spline	2.41±1.24 ($\underline{10^{-3}}$)	2.62±1.43 (0.71)	1.78±1.18 (0.53)	2.27±1.31 ($\underline{0.05}$)
Without Adaptation	2.78±1.82 ($\underline{10^{-5}}$)	3.82±2.57 ($\underline{10^{-6}}$)	2.30±1.66 ($\underline{10^{-4}}$)	2.96±2.44 ($\underline{10^{-6}}$)
Our method	**2.26±1.41**	**2.61±1.58**	**1.71±1.36**	**2.19±1.54**

4.4.3.3 Result

The loss curves showing the transition between the train and adaptation stages are shown in Fig. 4.8.

As shown in Fig. 4.9, due to the strong noise and artifacts, the classic method was mainly driven by the high-intensity gradient near the diaphragm region and failed to correctly estimate the motion within the lungs. Network without adaptation heavily underestimated the motion magnitude. The test-time adaptation generated a much more reasonable result with good intensity matching and feasible DVF.

Table 4.4: Target registration errors based on the landmarks on lung MRI. Results are provided as mean ± standard deviation (p-value from paired t-tests).

	TRE (mm)
Before	7.24±5.11
Classic B-spline	2.97±2.11 (10^{-3})
Without adaption	3.04±2.38 (10^{-5})
Our method	**2.00±1.63**

As shown in Table. 4.4, our method achieved the best TRE among the three, with

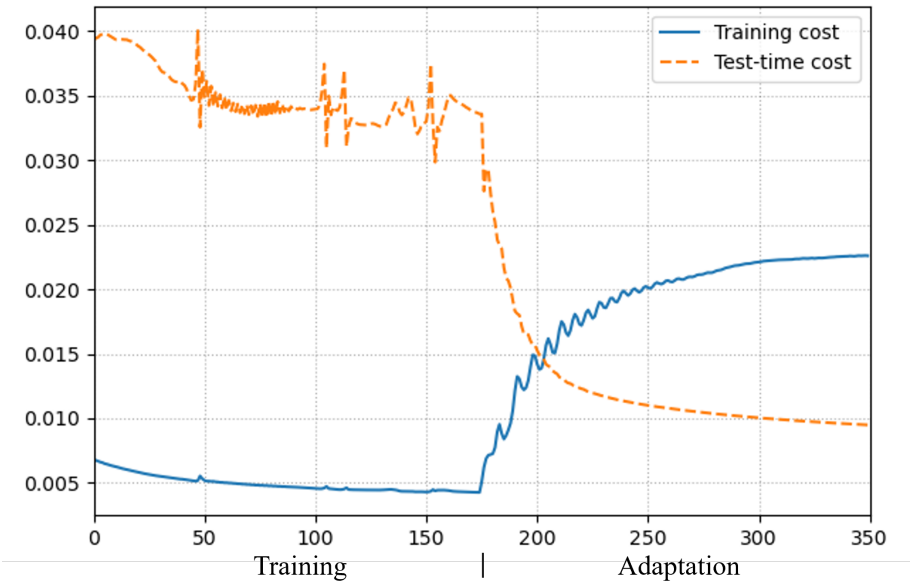

Figure 4.8: Loss curves for the cross-modality adaptation on lung MRI.

statistical significance ($p < 0.01$).

4.5 Discussion

The proposed test-time adaptation provides a logical synergy between the DL-based method and the classic optimization-based method: it combines the benefit of data-driven regularization and rich representation in the deep network modeling and general training process, and the focus on individual testing inputs from the adaptation stage, similar to the classic method.

While test-time adaptation yielded significant improvement in individualized registration results in all variations tested, its impact is related to the level or severity of domain shift. For moderate domain shift due to difference in imaging protocol and/or scanner setup, the cost functions for initial training and testing are consistent and the discrepancy can be largely considered as a consequence of the out-of-distribution behavior with the single sample, and sufficient adaptation can bring the test-time cost down to the level of training

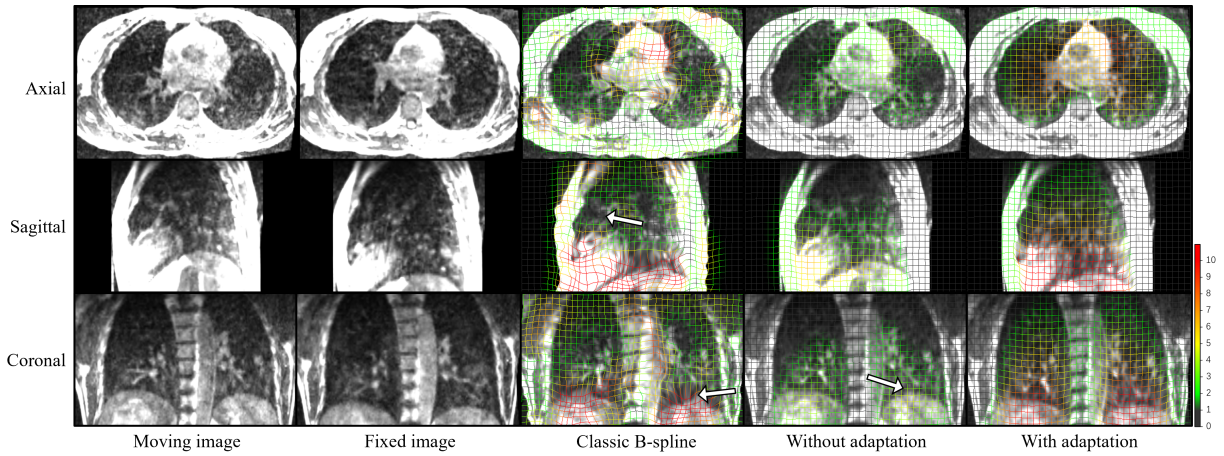

Figure 4.9: Example registration results on lung MRI.

cost at convergence. On the other hand, for more significant domain shift, due to platform or modality differences, as in the case of drastic MRI field strength change or CBCT vs MRI, the fidelity terms in the objective function are effectively measuring quite different quantities with their numerical values associated with different interpretations and possibly scales. In those cases, it is reasonable to expect that the test-time cost and training cost will improve and deteriorate during adaptation respectively. In addition, the convergence pattern and values during adaptation are expected to differ from the training stage as the network parameters are evolved with respect to a different objective.

While the test-time objective-based adaptation tailors the DIR network to the testing scenario, it is always important to harvest the strength of a good prior to regulate the objective and network model. From the perspective of addressing the probable discrepancy in noise and signal statistics between what the network was initially trained on and applied to, a stable prior compatible to both domains is desirable. In this study, we used a CAE-based motion feasibility descriptor with well-demonstrated behavior and strength [SCM21]. It meets the domain-invariant need, and is essential for driving the network update and imposing feasibility constraints during the test-time adaptation. The proposed adaptation scheme is also compatible with other prior generating schemes such as the adversarial training

or biomechanical models [HGG18].

We may expand the rationale of adaptation and individualization to a hierarchical scheme to facilitate clinical workflow. For example, a mature DIR network may be initially trained on the abundant 4D CBCT and adapted a patient's MRI images similar to Sec. 4.4.3. Then for longitudinal studies, a per-day adaptation can be further executed to boost the DIR performance for a specific testing session.

A limitation of the proposed test-time adaptation is the higher computation time compared to a typical evaluation-only DL inference module. As discussed previously, the burden of adaptation depends on various factors including the dissimilarity of the test sample to the training cohort and the level of domain shift. When neither of these is too large, semi-real-time execution is still possible, which suffices for a large portion of medical registration tasks. We are actively working on investigating a secondary DL approach to treat the adaptation as a different task and/or correction scheme to expedite the current approach.

4.6 Conclusion

We have proposed a test-time adaptation scheme to address the generalization gap between training and testing data and improve registration performance, by fitting the deep registration network to individual testing data. Extensive experiments showed that the test-time adaptation yielded significantly improved registration. We have demonstrated the potential to apply trained registration network to unseen data acquired on a different protocol or scanner, with a different imaging platform, or even modality.

CHAPTER 5

Continuous 4D Respiratory Motion Synthesis using a Conditional Registration Network

5.1 Introduction

Respiratory motion resolution is important in thoracic and abdominal imaging. 4DCT is a useful approach to characterize phase-resolved volumetric images. However, 4D data is typically acquired by assembling projections from multiple breathing cycles and sorting them according to pre-defined respiratory phases. Artifacts can occur if data segments were unbalanced, caused by device-dependent limitations of gantry rotation and the variability of the patient's respiratory pattern. Furthermore, 4DCT scan is more demanding in scanning time and is associated with higher radiation dose to the imaging subject, which can increase up to an order of magnitude [KSS04]. These factors together give rise to relatively sparse high-quality 4DCT data.

On the other hand, 4D information is critical to appreciate patient-specific dynamics, establish motion manifold, appreciate variations, and as important backbones to support important online effort such as real-time reconstruction [SZX19, MHS13].

Therefore, it is desirable to synthesize 4DCT from a protocol with a shorter time and lower dose demand. The requirement for such synthesis would be dependent on the end application, but should include agreement to high-quality 4DCT when available, nimbleness in phase definition, and efficiency. Recently, an effort has been made to utilize the pix2pix infrastructure to synthesize a prescribed set of phases [JV19].

In this study, we take a different perspective and approach the 4D problem from a dynamic imaging perspective. Realizing the intrinsic connections between various phase states in respiratory dynamics, we set out to model a spatiotemporally continuous deformation tensor manifold. A time/phase-specific snapshot of this manifold corresponds to a typical DVF from a reference coordinate. We consider a setup where the input is a pair of extreme respiratory states, i.e., EE and EI, and the output is the DVF (and correspondingly the volumetric image) for an arbitrary phase controlled by an input scalar parameter.

5.2 Method

5.2.1 4D Motion Synthesis

Fig. 5.1 shows the proposed method consisting of a conditional registration network and a spatial transformation module. The conditional registration network takes a conditional

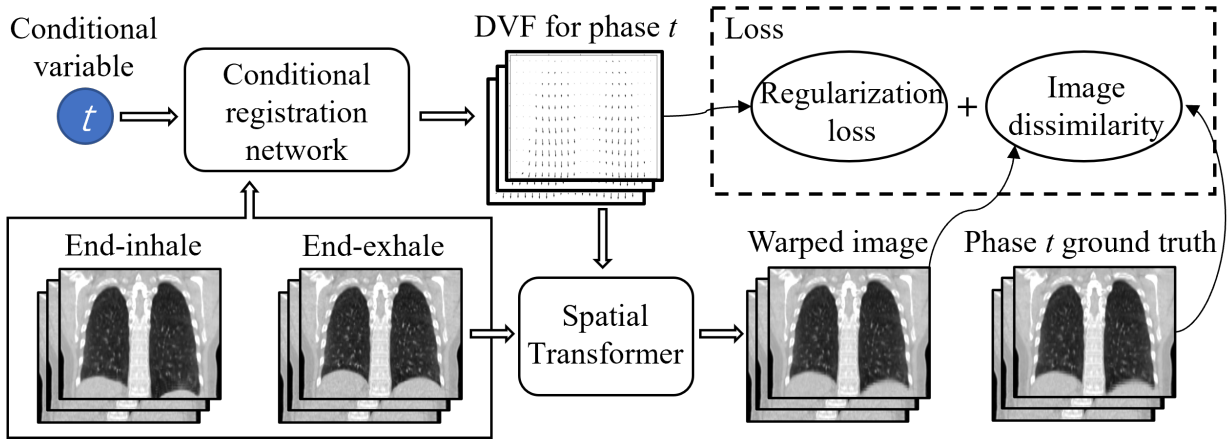

Figure 5.1: Overview of the proposed motion synthesis method. The conditional registration network takes the two extreme phases of a scan as input and outputs a DVF that correspond to the phase determined by the conditional variable t. The spatial transformer warps the end-exhale image with the DVF. Image dissimilarity to the phase-specific ground truth and DVF regularity are used to drive the network update.

variable t and concatenated EE-EI image pair (I_0, I_1) as input, and generates a DVF v_t, with which the EE image I_0 is warped toward an image I_t that corresponds to a certain phase denoted by t. Inside the spatial transformation module, a sampling grid is created using the DVF. The input moving image I_0 is sampled at these grid points to form the warped image $\hat{I}_t = I_0 \circ v_t$ [JSZ15, VBV19].

The loss function is defined as the weighted sum of an intensity match discrepancy and a regularity penalty:

$$\begin{aligned} L &= L_s(I_t, I_0 \circ v_t) + \lambda L_r(v_t) \\ &= L_s(I_t, I_0 \circ f(I_0, I_1, t; \theta)) + \lambda L_r(f(I_0, I_1, t; \theta)), \end{aligned} \tag{5.1}$$

where L_s is the image similarity loss, L_r is the DVF regularization loss, f is the function of the conditional registration network, θ is the network parameters, and λ is a balancing hyper-parameter. In this work, NCC is used as the image similarity metric L_s, and BP is used as the DVF regularization loss L_r to penalize non-smooth deformations and encourage physical feasibility [RSH99].

Upon review, a set of high-quality 4DCT scans are used as training samples. The conditional variable t is defined within the range [0,1], with $t = 0$ and $t = 1$ denoting the EE and EI phases, respectively. When $t = 0$, the network is supposed to output zero motion field, and when $t = 1$, it reduces to a conventional registration problem, with I_0 and I_1 being the moving and fixed images, respectively. A back-propagation scheme is used to derive the DVF solution to minimize the objective in Eq. (5.1) and update the network parameters θ. At test time, the trained network takes paired EE and EI images as input, and generates DVF and the corresponding image for any arbitrary phase by varying the conditional variable t.

5.2.2 Network Architecture

The proposed conditional registration network takes concatenated EE and EI images as input, and outputs a DVF conditioned on t. As shown in Fig. 5.2, the network uses a

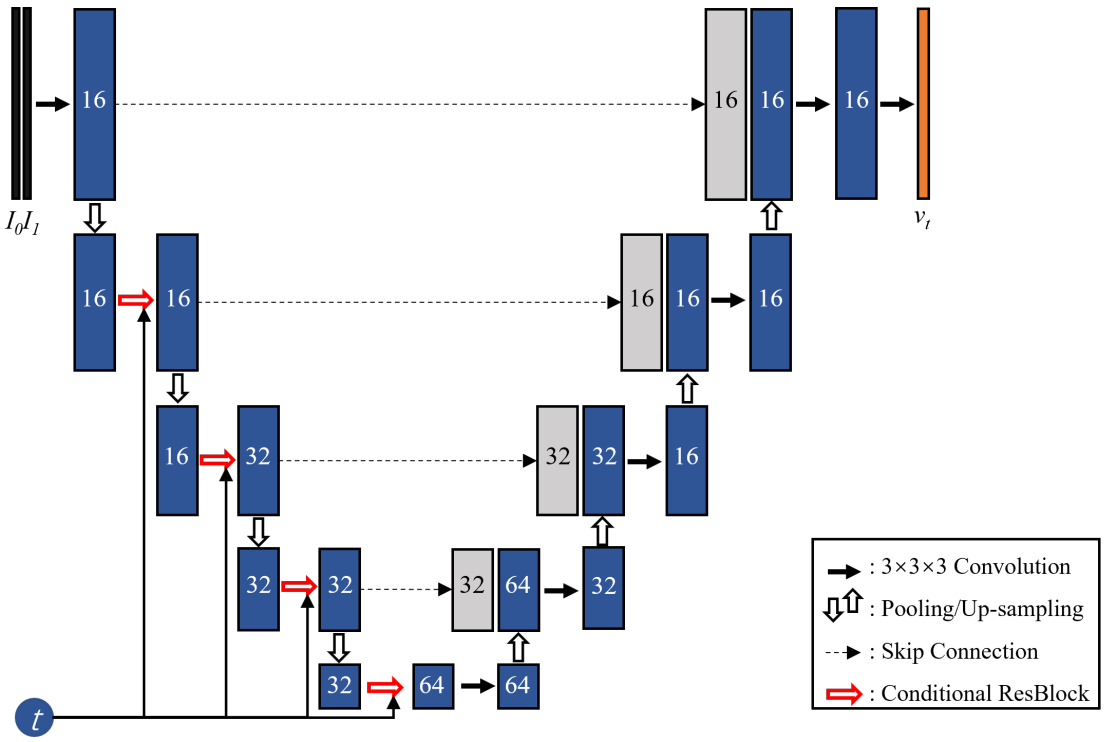

Figure 5.2: Architecture of the conditional registration network. Conditional residual blocks are used to replace the convolutions in the U-net encoding path and take a scalar parameter t as conditional variable input.

general U-net structure [RFB15] to take advantage of the hierarchical structure and skip connections for effective learning of features at all scales.

In order to condition the network model on the continuous breathing phase, we replace most of the convolution layer in the encoding path with a novel conditional residual block (CRB). As shown in Fig. 5.3, the CRB uses a fully-connected subnetwork to map the conditional variable t to two hidden variables a and b, which are then multiplied and added to the feature map within the residual block respectively to convey phase information [HZR16].

The fully-connected subnetwork has two layers, with 32 and 16 units, respectively, and uses ReLU activation. All the $3 \times 3 \times 3$ convolution layers use strides of one, zero-padding, and ReLU activation, except the last layer, which uses linear activation. Average pooling

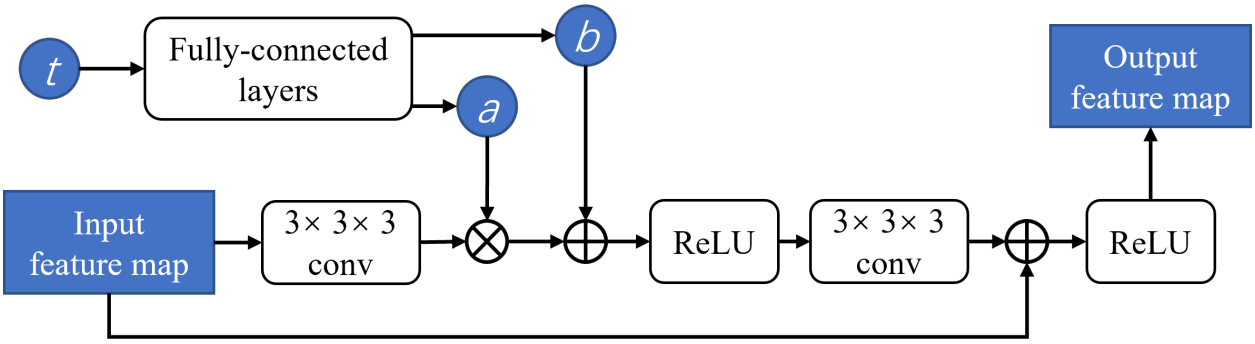

Figure 5.3: Architecture of the proposed conditional residual block. t is the input conditional variable. a and b are two scalar variables generated from the fully-connected subnetwork.

and up-sampling with a scaling factor of two are used in the encoding and decoding paths, respectively.

5.2.3 Calibration

In our training setup, the conditional variable has a direct correspondence with the phase in 4DCT. However, the regularization from the bending energy biases the network towards a slight under-estimation of the motion magnitude, resulting in a phase offset. To address this bias, we introduce a simple yet effective calibration to post-process the conditional registration network.

Specifically, an offset value t_0 is added to the conditional variable t before being fed into the network for inference.

$$v'_t = f(I_0, I_1, t'; \theta) = f(I_0, I_1, t + t_0; \theta). \tag{5.2}$$

The offset value is determined by back-searching for the optimal t' value that minimizes the MSE of image intensity on the training data (details in Sec. 5.3.6), and is fixed once the network training is complete. The linear relationship between t and the phase number persists.

5.3 Experiments and Results

5.3.1 Data

4DCT data from the 4D-Lung collection in the Cancer Imaging Archive (TCIA) [HWS17] was used in our experiment. They were acquired using a 16-slice helical CT scanner (Brilliance Big Bore, Philips Medical Systems) during the chemoradiotherapy of 20 locally advanced, non-small cell lung cancer patients. Each scan had 10 breathing phases, with 3 mm slice thickness. The in-plane resolution was 0.9766 mm for a grid size of 512×512.

All images were resampled to slice thickness 2.5 mm and in-plane pixel spacing 1.16 mm, and then cropped to a $256 \times 256 \times 96$ window that covered the lungs. Image intensities were clamped to a $[-1000, 500]$ HU window and normalized to a $[0, 1]$ range.

In this experiment, we focused on the inhalation dynamics and used six of the ten phases (0.0% to 50.0%) in training, with 0.0% being the EI phase (I_1) and 50.0% being the EE phase (I_0).

We used a cross-validation scheme and divided the 20 scans into four groups, each containing five scans. In the experiment, three of the groups were used as the training set and the remaining one was used for testing.

5.3.2 Implementation

The phase index t was assigned 0, 0.2, 0.4, 0.6, 0.8, 1.0, respectively, for the six-phase setup. The network was trained for 200 epochs with batch size 1. The balancing parameter $\lambda = 0.3$ was used for bending energy regularization. ADAM optimizer with learning rate 10^{-4} was used. The calibration offset value t_0 was estimated and set to 0.10.

5.3.3 Calibration Test

While a bias of phase correspondence is expected due to regularization, it is important to appreciate its severity and assess the impact or efficacy of the proposed calibration module.

To this end, we performed two sets of tests to characterize the upper and lower bound performance. The performance lower bound is established by executing the conditional registration network in the absence of the calibration module. This will quantify the impact of phase bias and also serves the role of ablation analysis for the calibration module. The performance upper bound is obtained by an Oracle study: for each phase in the test, we exhaustively searched for the t value that yielded the lowest MSE in image intensity. Note that this process is not only time-consuming but also reflects an impractical idealization, because ground truths for intermediate phases are unavailable during real inference.

5.3.4 Benchmark Methods

The proposed method was compared against two alternative benchmark approaches. Both methods were utilized on well-established pair-wise registration and used a linear scaling scheme to generate intermediate DVFs and the corresponding images. Comparison against these methods was expected to signify the impact of the proposed conditioning.

SE-Linear: Given an EE-EI image pair, a simple approach to synthesize motion field for any intermediate phase is to scale the EE-EI DVF. Specifically, classic B-spline method was used to generate the EE-EI DVF. Then, the DVF was multiplied by scaling factor $\alpha = 0.2, 0.4, 0.6, 0.8$ to obtain the intermediate phases. The classic B-spline registration was generated using the SimpleElastix toolbox [MBS16]. The same cost function as the unsupervised objective was used, and a multi-resolution strategy was adopted for optimization, with 30 optimization iterations in each of the four resolution levels.

DL-Linear: Similar to SE-Linear, we used the EE-EI DVF generated by the proposed conditional registration network, and applied linear scaling to obtain DVF for a target phase

and generated the corresponding image.

5.3.5 Evaluation

Without direct access to ground-truth 4D motion fields, the synthesis results are evaluated by image intensity. We quantified the performance with RMSE and SSIM to the known intermediate phase images in testing 4DCTs. Paired t-test was used to examine the statistical significance. Visualization was also generated to help appreciate and assess physiological feasibility qualitatively.

5.3.6 Results

5.3.6.1 Calibration Evaluation

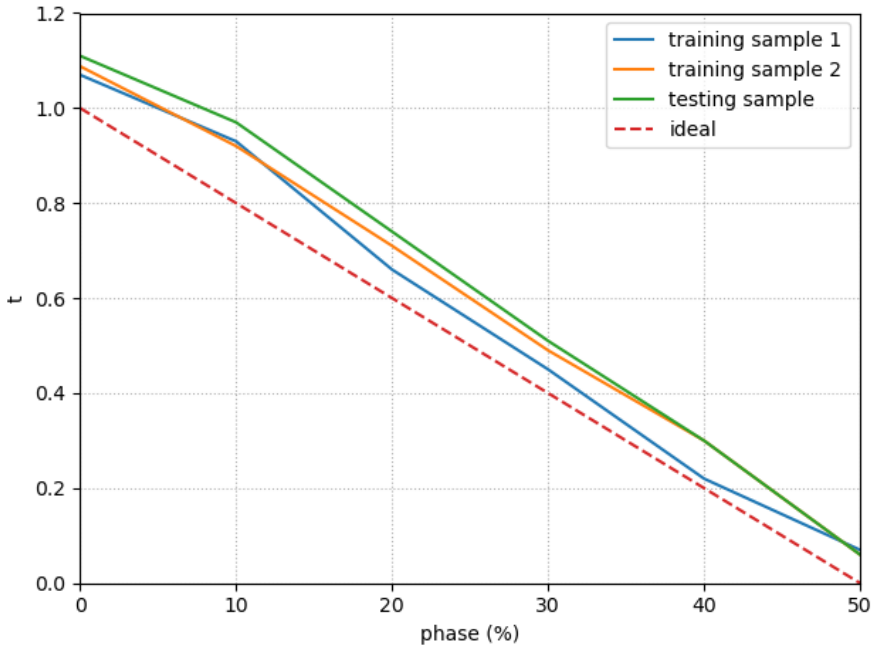

Figure 5.4: Example back-search results. The optimal t value has deviated from the "ideal" line, and the shift appears to be global.

As a part of the calibration process, for each phase of each data sample, we searched for the optimal t value to minimize the image intensity difference measured by MSE. Three example back-searching results are shown in Fig. 5.4. It can be observed that the optimal t value had deviated from the "ideal" line, but the linear relationship still persisted. The shift appeared to be global and applied for samples in either training and testing sets. The calibration process in Sec. 5.2.3 accommodated this shift. The offset value t_0 was set to be the mean deviation from the ideal line for all samples in the training set and was 0.10 in this experiment.

Table. 5.1 reports the impacts of phase bias and its correction. As expected, the oracle scheme based on exhaustive search yields the smallest RMSE and highest SSIM. Our proposed calibration achieved a similar SSIM to the Oracle correction, without statistically significant difference. Although the RMSE was statistically different from the Oracle correction, equivalence tests indicated that they were statistically equivalent ($p < 0.05$), in terms of either RMSE or SSIM.

Table 5.1: Impact of the calibration module. Results are provided as mean ± standard deviation (p-value from paired t-tests when comparing the method against ours).

	RMSE (HU)	SSIM
No Calibration	110.8±39.9 (10^{-5})	0.899±0.066 (10^{-5})
Oracle	**68.8±31.8** (10^{-3})	**0.927±0.040** (0.031)
Our Calibration	72.6±34.1	0.925±0.046

5.3.6.2 Efficacy of Conditioning

Table. 5.2 shows the quantitative results for model comparison between direct DVF scaling vs the proposed conditioning. Compared to the two benchmark methods, our method achieved much better results in terms of both RMSE and SSIM, with statistical significance. The average run time for generating a 10-phased 4D motion field and image was 2.85 s.

Table 5.2: Quantitative results on image synthesis. Results are provided as mean ± standard deviation (p-value from paired t-tests).

	RMSE (HU)	SSIM
SE-Linear	125.3±56.83 (10^{-6})	0.866±0.068 (10^{-5})
DL-Linear	138.6±59.72 (10^{-6})	0.859±0.079 (10^{-4})
Proposed Conditional DL	**72.6±34.1**	**0.925±0.046**

A qualitative comparison is shown in Fig. 5.5. As indicated by the red arrows, when synthesizing the phase at 30%, SE-Linear and DL-Linear severely underestimated the motion in the anterior and superior parts of the lungs. In addition, as indicated by the blue arrows, SE-Linear was strongly affected by the local artifacts and resulted in non-smooth DVFs. In comparison, the image generated by the proposed conditional DL method was closer to the ground truth. It correctly estimated the motion magnitude on both local and global levels, and was spatially smooth.

5.3.6.3 Qualitative Visualization for Arbitrary Continuous Phases

An example synthesis is shown in Fig. 5.6. Our method generated realistic DVFs that were spatiotemporally smooth. The method was able to synthesis any intermediate breathing phases, even for t values not included in the training setup (e.g., $t = 0.3$).

To further demonstrate the behavior of the proposed model, we calculated the motion magnitude at the lower lobe region, along the three imaging axes at different t values, and The relative motion magnitude change calculated by dividing the magnitude with the directional maximum that correspond to EI and is shown in Fig. 5.7. Our model was able to capture the temporally non-linearity of the respiratory motion.

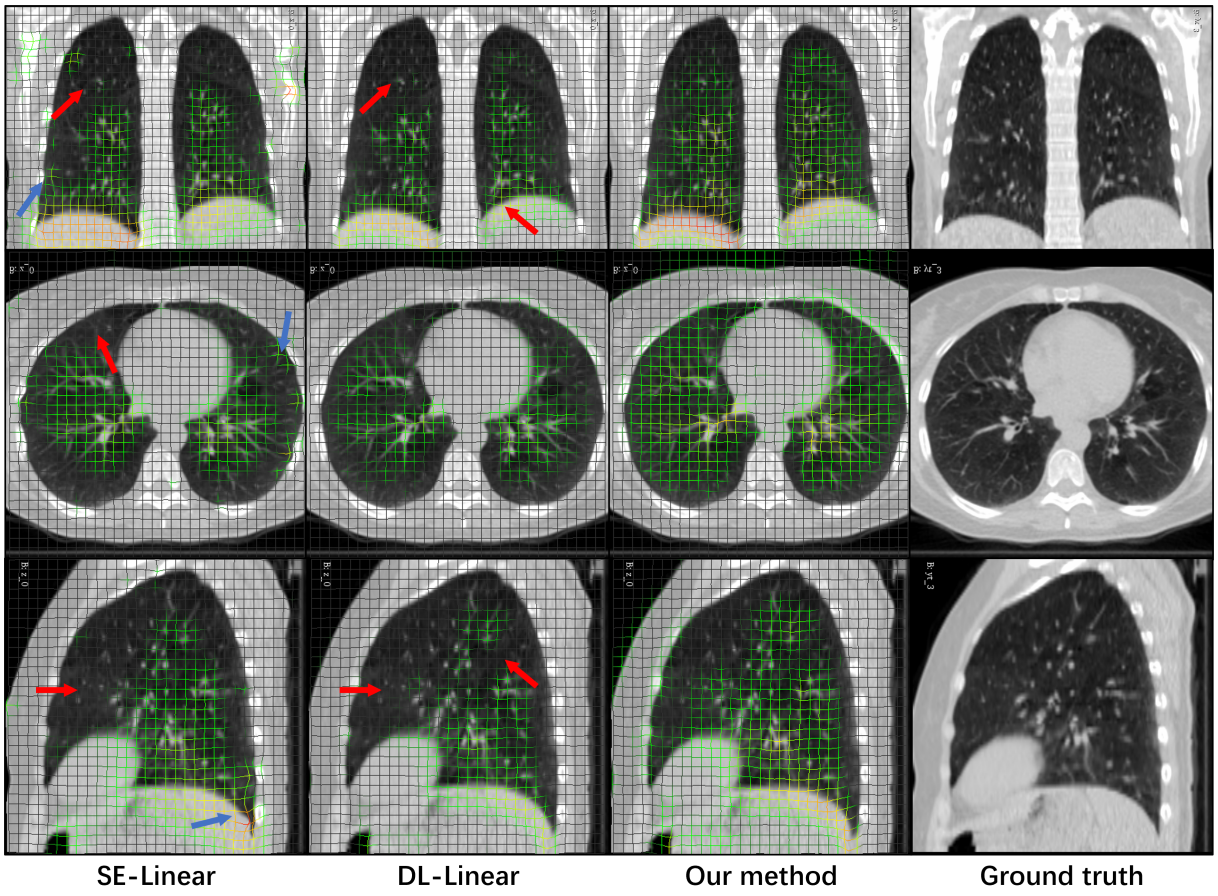

Figure 5.5: Example synthesis results for phase 30%, using the benchmark methods and the proposed approach. Red arrows indicate severe underestimation of motion magnitude. Blue arrows indicate significant local non-smoothness.

5.4 Discussion

In this study, we used the weighted sum of NCC and bending energy penalty as the loss function for the network training. The proposed paradigm is compatible with any differentiable similarity metric and regularizer. We have shown that the proposed calibration module, albeit simple, can correct for regularization-induced phase bias effectively.

In contrast to an existing 4DCT synthesis method [JV19] where parallel image-to-image translation networks to map a static CT to multiple predetermined breathing phases, our

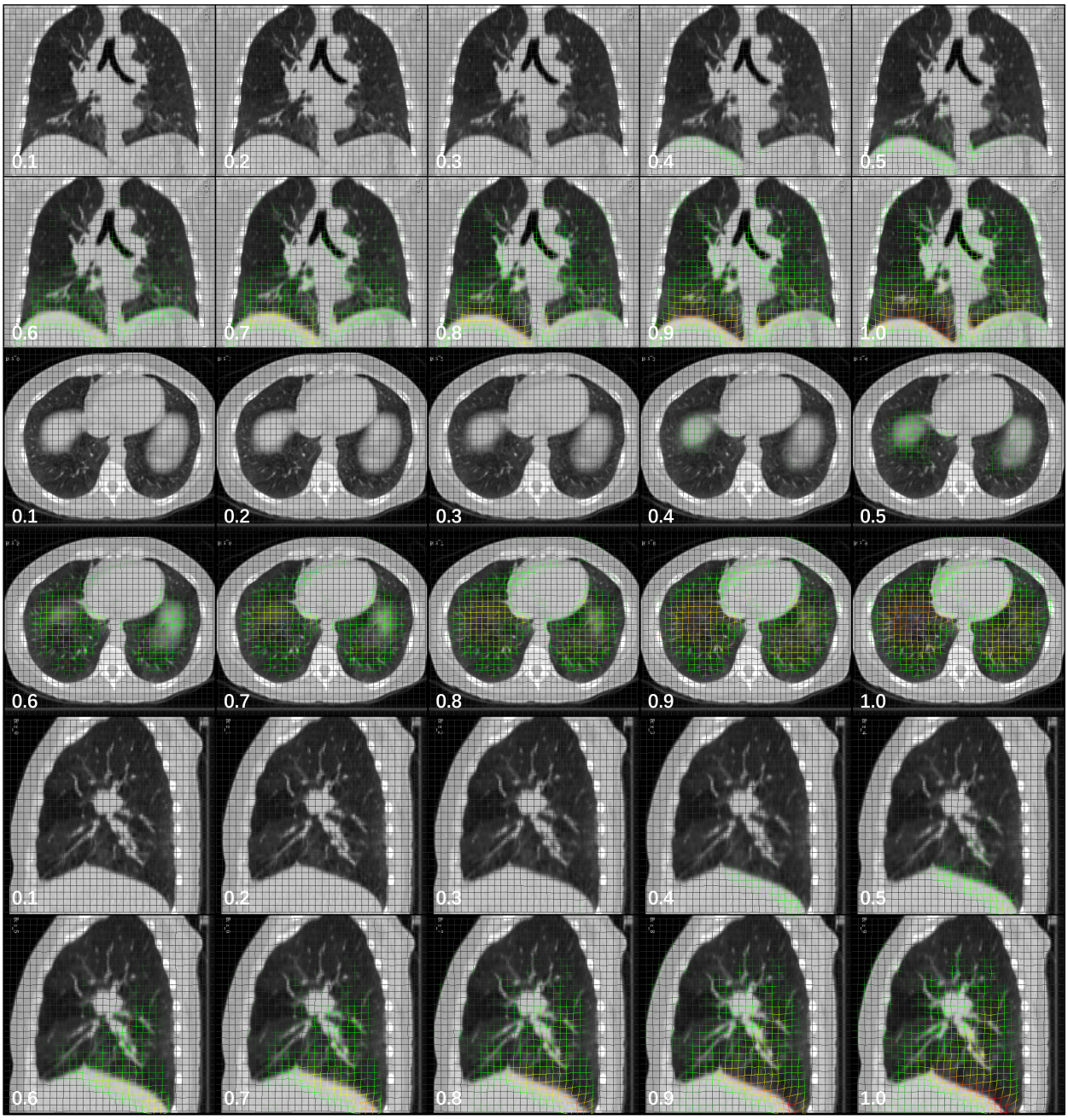

Figure 5.6: Example synthesis results. The numbers indicate the t values.

method chooses to centralize on DVF modeling and then translates it to image generation. In doing so, it manages to capture the coupling across different phases and incorporates temporal consistency naturally. By modeling the underlying 4D motion with a continuous

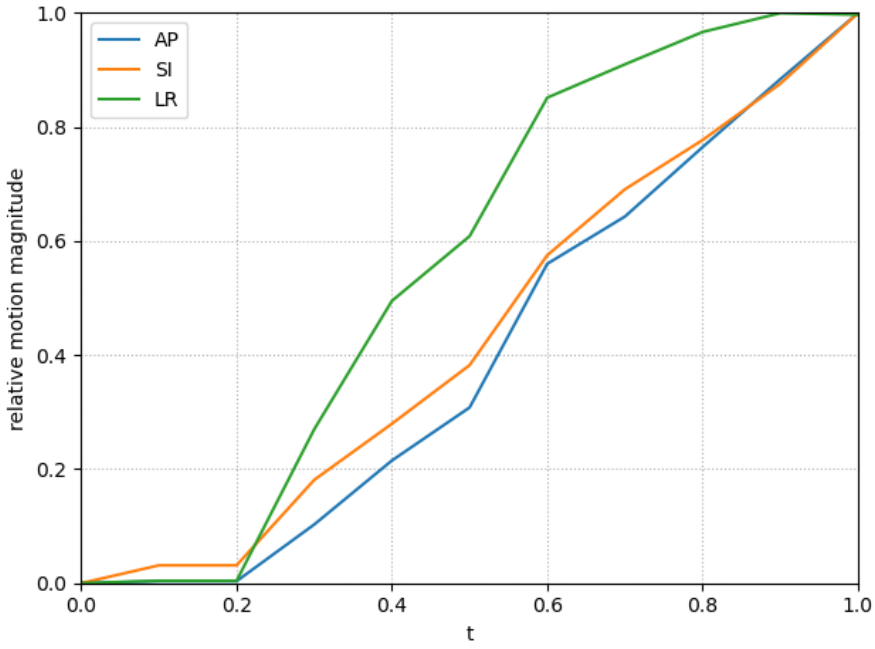

Figure 5.7: Curves showing the change of relative motion magnitude along different axes. The average within a region located at the lower lobes was used, and the magnitude was normalized with respect to directional maximum for display. The normalization factors were 4.1, 6.8, and 2.9 mm for the AP, SI, LR respectively.

conditional variable, our model can generate arbitrary intermediate phases on the fly, without the restriction of pre-set phase banks. Finally, with the synthesis anchored at extreme stage pairs, patient-specific motion information, including but not limited to respiratory volume, is incorporated better. This study also evaluated the synthesis comprehensively based on the complete intensity profile, beyond surrogate metrics such as lung volume in [JV19].

One limitation of the current synthesis is its suboptimal performance in intensity-homogeneous regions outside the lungs due to the lack of local structural contrast. In future work, we plan to incorporate respiratory motion-specific feasibility priors from either registration or biomechanical models [SR21, HGG18]. We expect such prior to perform better than prescribed universal bending energy to help yield more realistic motion synthesis. Alleviated bias from

better prior may also help to make the conditional variable more congruent with the phase number, eliminating the need for explicit calibration.

5.5 Conclusion

We have presented a novel paradigm to synthesize continuous 4D respiratory motion from EI and EE image pairs. Our method trains a conditional image registration network with 4DCT data. By varying the conditional variable, the network can generate DVF for an arbitrary intermediate breathing phase. Experiments have demonstrated the method's ability to precisely control the breathing phase, and generate spatiotemporally smooth realistic DVFs. This method can be applied in common clinical practice to support 4DCT synthesis, online reconstruction, and other downstream tasks.

CHAPTER 6

Summary

In this dissertation, we have developed modules and methods for deformable image registration by combining physical-driven rationales with advanced learning techniques. Our methods effectively addressed several major challenges in DL-DIR, leading to improved registration efficiency, accuracy, and robustness.

In Chapter 2, we have proposed a DIR network that is conscious of and self-adaptive to deformation of various scales to improve registration performance. The introduction and integration of dilated inception modules and scale adaptation modules address the heterogeneous scale problem with self-adaptation and high efficiency. The method serves as an efficacious alternative to the time-consuming multi-resolution strategy.

In Chapter 3, we have proposed two different approaches to incorporate learned implicit feasibility conditions into DIR. In the supervised method, a Plug-and-Play feasible motion prior is developed from high-quality images, and then incorporated as a regularizer to train an unsupervised DIR network. In the unsupervised method, novel deformation parametrization in the form of a generator network is developed to learn implicit feasibility conditions on DVF from paired images using the alternating back-propagation algorithm. Both methods managed to model feasibility conditions without external regularization and led to physically and physiologically more reasonable DVFs.

In Chapter 4, we have made investigations to address the potential domain shift and to improve the accuracy and robustness of registration. A test-time adaptation scheme is used to fit trained DIR networks to individual testing data. The method yielded significantly

improved registration on unseen data in cross-protocol, cross-platform, and cross-modality scenarios.

In Chapter 5, we have proposed a DIR approach to synthesize continuous 4D motion from image pairs at two phases. The method trains a conditional image registration network with 4DCT data. By varying the conditional variable, the network can generate DVF for an arbitrary intermediate breathing phase. This novel paradigm can be applied to augment the available data or serve as a backbone to support online reconstruction and downstream analysis from real-time imaging.

Each of these developments not only bridges the current gap of existing methods to clinical needs, but also alludes to further investigation directions.

From the technical perspective, the current supervised CAE based prior was trained with DVFs derived from numerical registration on high-quality images. While this approach is robust enough to generate physiologically reasonable breathing motion fields from lung CTs, it is more challenging to generate DVFs that faithfully represent cardiac motions. It is worthwhile to complement this effort with alternatives, such as dynamic anthropomorphic phantoms or biomechanical models, to overcome the limitations of image quality and registration algorithms.

The feasibility conditions on DVF learned from the unsupervised generator network can be furthered by (1) going from paired setting to extended 4D temporal domain as well; and (2) using a dual generator network to disentangle the anatomy and deformation information from a collection of multiple-subject multiple-instance images.

It is also expected that combining registration with other image processing tasks may be beneficial in enhancing the connection to downstream clinical endpoints and complement the information extraction during the learning process. While preliminary assessment has been performed in Section 3.2.2 for motion-compensated image enhancement, integration of the DIR to reconstruction is a natural next step. We also plan to develop a registration-

segmentation multi-task network to improve the network's natural attention to important singularities or critical structures.

From the application perspective, our current experiments mainly focused on inter-phase DIR of 4D heart and lung images, where the estimated DVF represents intra-subject cardiac or respiratory motions. The methods proposed in Chapters 2-4 can be applied to other scenarios such as inter-subject, inter-modality, or longitudinal DIR where tissue correspondence needs to be established.

REFERENCES

[Ash07] John Ashburner. "A fast diffeomorphic image registration algorithm." *Neuroimage*, **38**(1):95–113, 2007.

[ASS17] Ehsan Abadi, William P Segars, Gregory M Sturgeon, Justus E Roos, Carl E Ravin, and Ehsan Samei. "Modeling lung architecture in the XCAT series of phantoms: Physiologically based airways, arteries and veins." *IEEE Trans. Med. Imag.*, **37**(3):693–702, 2017.

[BDH16] Ander Biguri, Manjit Dosanjh, Steven Hancock, and Manuchehr Soleimani. "TIGRE: a MATLAB-GPU toolbox for CBCT image reconstruction." *Biomedical Physics & Engineering Express*, **2**(5):055010, 2016.

[BEK19] Riddhish Bhalodia, Shireen Y Elhabian, Ladislav Kavan, and Ross T Whitaker. "A cooperative autoencoder for population-based regularization of CNN image registration." In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, pp. 391–400. Springer, 2019.

[BZS19] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. "VoxelMorph: a learning framework for deformable medical image registration." *IEEE Trans. Med. Imag.*, 2019.

[CBH02] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *J. Artif. Intell. Res.*, **16**:321–357, 2002.

[CCG09] Richard Castillo, Edward Castillo, Rudy Guerra, Valen E Johnson, Travis McPhail, Amit K Garg, and Thomas Guerrero. "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets." *Phys. Med. Biol.*, **54**(7):1849, 2009.

[CCM09] Edward Castillo, Richard Castillo, Josue Martinez, Maithili Shenoy, and Thomas Guerrero. "Four-dimensional deformable image registration using trajectory modeling." *Phys. Med. Biol.*, **55**(1):305, 2009.

[CDD10] Kunlin Cao, Kaifang Du, Kai Ding, Joseph M Reinhardt, and Gary E Christensen. "Regularized nonrigid registration of lung CT images by preserving tissue volume and vesselness measure." *Grand Challenges in Medical Image Analysis*, pp. 43–54, 2010.

[CGW19] Zhangpei Cheng, Kaixuan Guo, Changfeng Wu, Jiankun Shen, and Lei Qu. "U-Net cascaded with dilated convolution for medical image registration." In *2019 Chinese Automation Congress*, pp. 3647–3651. IEEE, 2019.

[CPK17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." *IEEE Trans. Pattern Anal. Mach. Intell.*, **40**(4):834–848, 2017.

[CYZ18] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Qian Wang, Pew-Thian Yap, and Dinggang Shen. "Deformable image registration using a cue-aware deep regression network." *IEEE Trans. Biomed. Eng.*, **65**(9):1900–1911, 2018.

[DBG18] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. "Unsupervised learning for fast probabilistic diffeomorphic registration." In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, pp. 729–738, 2018.

[DBG19] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces." *Med. Image Anal.*, **57**:226–236, 2019.

[Dic45] Lee R Dice. "Measures of the amount of ecologic association between species." *Ecology*, **26**(3):297–302, 1945.

[ELV19] Koen AJ Eppenhof, Maxime W Lafarge, Mitko Veta, and Josien PW Pluim. "Progressively trained convolutional neural networks for deformable image registration." *IEEE Trans. Med. Imag.*, **39**(5):1594–1604, 2019.

[EP18] Koen AJ Eppenhof and Josien PW Pluim. "Pulmonary CT registration through supervised learning with convolutional neural networks." *IEEE Trans. Med. Imag.*, **38**(5):1097–1105, 2018.

[FB20] Tobias Fechter and Dimos Baltas. "One-Shot Learning for Deformable Medical Image Registration and Periodic Motion Tracking." *IEEE Trans. Med. Imag.*, **39**(7):2506–2517, 2020.

[FCW19] Jingfan Fan, Xiaohuan Cao, Qian Wang, Pew-Thian Yap, and Dinggang Shen. "Adversarial learning for mono-or multi-modal registration." *Med. Image Anal.*, **58**:101545, 2019.

[FDK84] Lee A Feldkamp, Lloyd C Davis, and James W Kress. "Practical cone-beam algorithm." *Josa a*, **1**(6):612–619, 1984.

[FLW20a] Yabo Fu, Yang Lei, Tonghe Wang, Kristin Higgins, Jeffrey D Bradley, Walter J Curran, Tian Liu, and Xiaofeng Yang. "LungRegNet: An unsupervised deformable image registration method for 4D-CT lung." *Med. Phys.*, **47**(4):1763–1774, 2020.

[FLW20b] Yabo Fu, Yang Lei, Tonghe Wang, Kristin Higgins, Jeffrey D Bradley, Walter J Curran, Tian Liu, and Xiaofeng Yang. "An unsupervised deep learning approach for 4DCT lung deformable image registration." In *Proc. SPIE Med. Imag.*, volume 11313, p. 113132T, 2020.

[FWT19] Yabo Fu, Xue Wu, Allan M Thomas, Harold H Li, and Deshan Yang. "Automatic large quantity landmark pairs detection in 4DCT lung images." *Med. Phys.*, **46**(10):4490–4501, 2019.

[GCY20] Xiaoqing Guo, Zhen Chen, and Yixuan Yuan. "Complementary network with adaptive receptive fields for melanoma segmentation." In *Proc. IEEE 17th Int. Symp. Biomed. Imaging*, pp. 2010–2013, 2020.

[GKS16] Frank Godenschweger, Urte Kägebein, Daniel Stucht, Uten Yarach, Alessandro Sciarra, Renat Yakupov, Falk Lüsebrink, Peter Schulze, and Oliver Speck. "Motion correction in MRI of the brain." *Phys. Med. Biol.*, **61**(5):R32, 2016.

[GKZ15] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. "Domain generalization for object recognition with multi-task autoencoders." In *Proc. IEEE Int. Conf. Comput. Vis*, pp. 2551–2559, 2015.

[HBG21] Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. "SynthMorph: learning contrast-invariant registration without acquired images." *IEEE Trans. Med. Imag.*, 2021.

[HG09] Haibo He and Edwardo A Garcia. "Learning from imbalanced data." *IEEE Trans. Knowl. Data Eng.*, **21**(9):1263–1284, 2009.

[HGG18] Yipeng Hu, Eli Gibson, Nooshin Ghavami, Ester Bonmati, Caroline M Moore, Mark Emberton, Tom Vercauteren, J Alison Noble, and Dean C Barratt. "Adversarial Deformation Regularization for Training Image Registration Neural Networks." In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, pp. 774–782, 2018.

[HJB10] Mattias Paul Heinrich, Mark Jenkinson, Michael Brady, and Julia Schnabel. "Discontinuity preserving regularisation for variational optical-flow registration using the modified Lp norm." In *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 185–194, 2010.

[HLL16] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. "Learning deep representation for imbalanced classification." In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 5375–5384, 2016.

[HLZ17] Tian Han, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. "Alternating back-propagation for generator network." In *Proc. 31st AAAI Conf. Artif. Intell.*, 2017.

[Hol07] Mark Holden. "A review of geometric transformations for nonrigid body registration." *IEEE Trans. Med. Imag.*, **27**(1):111–128, 2007.

[HWS17] Geoffrey D Hugo, Elisabeth Weiss, William C Sleeman, Salim Balik, Paul J Keall, Jun Lu, and Jeffrey F Williamson. "A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer." *Med. Phys.*, **44**(2):762–771, 2017.

[HZR16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.

[Jap00] Nathalie Japkowicz. "The class imbalance problem: Significance and strategies." In *Proc. Int. Conf. on Artif. Intell.*, volume 56. Citeseer, 2000.

[JC02] Hans J Johnson and Gary E Christensen. "Consistent landmark and intensity-based image registration." *IEEE Trans. Med. Imag.*, **21**(5):450–461, 2002.

[JHG19] Amod Jog, Andrew Hoopes, Douglas N Greve, Koen Van Leemput, and Bruce Fischl. "PSACNN: Pulse sequence adaptive fast whole brain segmentation." *NeuroImage*, **199**:553–569, 2019.

[JLZ21] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zhibo Chen. "CASINet: Content-Adaptive Scale Interaction Networks for scene parsing." *Neurocomputing*, **419**:9 – 22, 2021.

[JSW02] David A Jaffray, Jeffrey H Siewerdsen, John W Wong, and Alvaro A Martinez. "Flat-panel cone-beam computed tomography for image-guided radiation therapy." *Int. J. Radiat. Oncol. Biol. Phys.*, **53**(5):1337–1349, 2002.

[JSZ15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. "Spatial transformer networks." In *Adv. Neural Inf. Process. Syst.*, pp. 2017–2025, 2015.

[JV19] Vincent Jaouen and Dimitris Visvikis. "4D respiratory motion synchronized image synthesis from static CT images using GANs." In *Proc. IEEE Nucl. Sci. Symp. Med. Imaging Conf.*, 2019.

[JYG20] Zhuoran Jiang, Fang-Fang Yin, Yun Ge, and Lei Ren. "A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration." *Phys. Med. Biol.*, **65**(1):015011, 2020.

[JYK12] Zhi-Qin Jiang, Kunyu Yang, Ritsuko Komaki, Xiong Wei, Susan L Tucker, Yan Zhuang, Mary K Martel, Sastray Vedam, Peter Balter, Guangying Zhu, et al. "Long-term clinical outcome of intensity-modulated radiotherapy for inoperable non-small cell lung cancer: the MD Anderson experience." *Int. J. Radiat. Oncol. Biol. Phys.*, **83**(1):332–339, 2012.

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv:1412.6980*, 2014.

[KEC21] Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu. "Test-time adaptable neural networks for robust medical image segmentation." *Med. Image Anal.*, **68**:101907, 2021.

[KMD17] Julian Krebs, Tommaso Mansi, Hervé Delingette, Li Zhang, Florin C Ghesu, Shun Miao, Andreas K Maier, Nicholas Ayache, Rui Liao, and Ali Kamen. "Robust non-rigid registration through agent-based action learning." In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, pp. 344–352, 2017.

[KSS04] Paul J Keall, George Starkschall, HEE Shukla, Kenneth M Forster, Vivian Ortiz, CW Stevens, Sastry S Vedam, Rohini George, Thomas Guerrero, and Radhe Mohan. "Acquiring 4D thoracic CT scans using a multislice helical method." *Phys. Med. Biol.*, **49**(10):2053, 2004.

[LF18] Hongming Li and Yong Fan. "Non-rigid image registration using self-supervised fully convolutional networks without training data." In *Proc. IEEE 15th Int. Symp. Biomed. Imaging*, pp. 1075–1078, 2018.

[LRR08] Jan JW Lagendijk, Bas W Raaymakers, Alexander JE Raaijmakers, Johan Overweg, Kevin J Brown, Ellen M Kerkhof, Richard W van der Put, Björn Hårdemark, Marco van Vulpen, and Uulke A van der Heide. "MRI/linac integration." *Radiother. Oncol.*, **86**(1):25–29, 2008.

[LYG19] Jingcong Li, Zhu Liang Yu, Zhenghui Gu, Hui Liu, and Yuanqing Li. "Dilated-inception net: multi-scale feature aggregation for cardiac right ventricle segmentation." *IEEE Trans. Biomed. Eng.*, **66**(12):3499–3508, 2019.

[MAS18] Dwarikanath Mahapatra, Bhavna Antony, Suman Sedai, and Rahil Garnavi. "Deformable medical image registration using generative adversarial networks." In *Proc. IEEE 15th Int. Symp. Biomed. Imaging*, pp. 1449–1453, 2018.

[MB06] R Mohan and T Bortfeld. "The potential and limitations of IMRT: a physicist's point of view." In *Image-Guided IMRT*, pp. 11–18. Springer, 2006.

[MBS16] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. "SimpleElastix: A user-friendly, multi-lingual library for medical image registration." In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 134–142, 2016.

[MHS13] Jamie R McClelland, David J Hawkes, Tobias Schaeffter, and Andrew P King. "Respiratory motion models: a review." *Med. Image Anal.*, **17**(1):19–42, 2013.

[MNJ20] M McNitt-Gray, S Napel, A Jaggi, SA Mattonen, L Hadjiiski, M Muzi, D Goldgof, Y Balagurunathan, LA Pierce, PE Kinahan, et al. "Standardization in quantitative imaging: a multicenter comparison of radiomic features from different software packages on digital reference objects and patient data sets." *Tomography*, **6**(2):118–128, 2020.

[MPA17] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. "Unified deep supervised domain adaptation and generalization." In *Proc. IEEE Int. Conf. Comput. Vis*, pp. 5715–5725, 2017.

[OT14] Francisco PM Oliveira and Joao Manuel RS Tavares. "Medical image registration: a review." *Comput. Methods. Biomech. Biomed. Engin.*, **17**(2):73–93, 2014.

[PHF14] Bartłomiej W Papież, Mattias P Heinrich, Jérome Fehrenbach, Laurent Risser, and Julia A Schnabel. "An implicit sliding-motion preserving regularisation via bilateral filtering for deformable image registration." *Med. Image Anal.*, **18**(8):1299–1311, 2014.

[PKH20] Bum Woo Park, Jeong Kon Kim, Changhoe Heo, and Kye Jin Park. "Reliability of CT radiomic features reflecting tumour heterogeneity according to image quality and image processing parameters." *Sci. Rep.*, **10**(1):1–13, 2020.

[PST13] David A Palma, Suresh Senan, Kayoko Tsujino, Robert B Barriger, Ramesh Rengan, Marta Moreno, Jeffrey D Bradley, Tae Hyun Kim, Sara Ramella, Lawrence B Marks, et al. "Predicting radiation pneumonitis after chemoradiation therapy for lung cancer: an international individual patient data meta-analysis." *Int. J. Radiat. Oncol. Biol. Phys.*, **85**(2):444–450, 2013.

[RDH17] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. "SVF-Net: Learning deformable image registration using shape matching." In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, pp. 266–274, 2017.

[REF09] Dan Ruan, Selim Esedoglu, and Jeffrey A Fessler. "Discriminative sliding preserving regularization in medical image registration." In *Proc. IEEE Int. Symp. Biomed. Imaging*, pp. 430–433, 2009.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, pp. 234–241, 2015.

[RFR06] Dan Ruan, Jeffrey A Fessler, Michael Roberson, James Balter, and Marc Kessler. "Nonrigid registration using regularization that accommodates local tissue rigidity." In *Proc. SPIE Med. Imag.*, volume 6144, p. 614412, 2006.

[RLC09] P Radau, Y Lu, K Connelly, G Paul, A Dick, and G Wright. "Evaluation framework for algorithms segmenting short axis cardiac MRI." *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, **49**, 2009.

[RSH99] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. "Nonrigid registration using free-form deformations: application to breast MR images." *IEEE Trans. Med. Imag.*, **18**(8):712–721, 1999.

[SCM21] Yudi Sang, Minsong Cao, Michael McNitt-Gray, Yu Gao, Peng Hu, Ran Yan, Yingli Yang, and Dan Ruan. "Enhancing 4D Cardiac MRI Registration Network with a Motion Prior Learned From Coronary CTA." In *Proc. IEEE Int. Symp. Biomed. Imaging*, pp. 917–920. IEEE, 2021.

[SCX18] Chang Shu, Xi Chen, Qiwei Xie, and Hua Han. "An unsupervised network for fast microscopic image registration." In *Proc. SPIE Med. Imag.*, volume 10581, p. 105811D, 2018.

[SDP13] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. "Deformable medical image registration: a survey." *IEEE Trans. Biomed. Eng.*, **32**(7):1153, 2013.

[SGK20] LI Song, KF Geoffrey, and HE Kaijian. "Bottleneck feature supervised U-Net for pixel-wise liver and tumor segmentation." *Expert Syst. Appl.*, **145**:113131, 2020.

[SGL19] Chun-Chien Shieh, Yesenia Gonzalez, Bin Li, Xun Jia, Simon Rit, Cyril Mory, Matthew Riblett, Geoffrey Hugo, Yawei Zhang, Zhuoran Jiang, et al. "SPARE: Sparse-view reconstruction challenge for 4D cone-beam CT from a 1-min scan." *Med. Phys.*, **46**(9):3799–3811, 2019.

[SIV17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. "Inception-v4, inception-resnet and the impact of residual connections on learning." In *Proc. 31st AAAI Conf. Artif. Intell.*, 2017.

[SJZ17] Wuzhen Shi, Feng Jiang, and Debin Zhao. "Single image super-resolution with dilated convolution based multi-scale information learning inception module." In *Proc. IEEE Int. Conf. Image Process.*, pp. 977–981. IEEE, 2017.

[SLJ15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1–9, 2015.

[SPC18] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. "Generalizing across domains via cross-gradient training." *arXiv preprint arXiv:1804.10745*, 2018.

[SR20] Yudi Sang and Dan Ruan. "Enhanced Image Registration With a Network Paradigm and Incorporation of a Deformation Representation Model." In *Proc. IEEE 17th Int. Symp. Biomed. Imaging*, pp. 91–94, 2020.

[SR21] Yudi Sang and Dan Ruan. "4D-CBCT Registration with a FBCT-derived Plug-and-Play Feasibility Regularizer." In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*. Springer, 2021.

[SSM10] WP Segars, G Sturgeon, S Mendonca, Jason Grimes, and Benjamin MW Tsui. "4D XCAT phantom for multimodality imaging research." *Med. Phys.*, **37**(9):4902–4915, 2010.

[SVB17] Hessam Sokooti, Bob de Vos, Floris Berendsen, Boudewijn PF Lelieveldt, Ivana Išgum, and Marius Staring. "Nonrigid image registration using multi-scale 3D convolutional neural networks." In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, pp. 232–239, 2017.

[SVI16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2818–2826, 2016.

[SWL20] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. "Test-time training with self-supervision for generalization under distribution shifts." In *International Conference on Machine Learning*, pp. 9229–9248. PMLR, 2020.

[SZX19] Liyue Shen, Wei Zhao, and Lei Xing. "Patient-Specific Reconstruction of Volumetric Computed Tomography Images from a Single Projection View via Deep Learning." *Nature Biomedical Engineering*, **3**:880–8, 2019.

[TDM13] Catalina Tobon-Gomez, Mathieu De Craene, Kristin Mcleod, Lennart Tautz, Wenzhe Shi, Anja Hennemuth, Adityo Prakosa, Hengui Wang, Gerry Carr-White, Stam Kapetanakis, et al. "Benchmarking framework for myocardial tracking and deformation algorithms: An open access database." *Med. Image Anal.*, **17**(6):632–648, 2013.

[UWS10] Martin Urschler, Manuel Werlberger, Eva Scheurer, and Horst Bischof. "Robust optical flow based deformable registration of thoracic CT images." *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 195–204, 2010.

[VBV17] Bob D de Vos, Floris F Berendsen, Max A Viergever, Marius Staring, and Ivana Išgum. "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*, pp. 204–212. Springer, 2017.

[VBV19] Bob D de Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. "A deep learning framework for unsupervised affine and deformable image registration." *Med. Image Anal.*, **52**:128–143, 2019.

[VPP09] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. "Diffeomorphic demons: Efficient non-parametric image registration." *NeuroImage*, **45**(1):S61–S72, 2009.

[VQL20] Andreas Venzke, Guannan Qu, Steven Low, and Spyros Chatzivasileiadis. "Learning optimal power flow: Worst-case guarantees for neural networks." In *IEEE Int. Conf. Commun. Control Comput. Technol. Smart Grids*, pp. 1–7. IEEE, 2020.

[VVS20] Bob D de Vos, Bas HM van der Velden, Jörg Sander, Kenneth GA Gilhuijs, Marius Staring, and Ivana Išgum. "Mutual information for unsupervised deep learning image registration." In *Proc. SPIE Med. Imag.*, volume 11313, p. 113130R, 2020.

[WAH20] Dongming Wei, Sahar Ahmad, Yunzhi Huang, Lei Ma, Qian Wang, Pew-Thian Yap, and Dinggang Shen. "An Auto-Context Deformable Registration Network for Infant Brain MRI." *arXiv:2005.09230*, 2020.

[WKL21] Dufan Wu, Kyungsang Kim, and Quanzheng Li. "Low-dose CT reconstruction with Noise2Noise network and testing-time fine-tuning." *Med. Phys.*, 2021.

[WL16] Jürgen Weese and Cristian Lorenz. "Four challenges in medical image analysis from an industrial perspective.", 2016.

[WLT19] Bo Wang, Yang Lei, Sibo Tian, Tonghe Wang, Yingzi Liu, Pretesh Patel, Ashesh B Jani, Hui Mao, Walter J Curran, Tian Liu, et al. "Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation." *Med. Phys.*, **46**(4):1707–1718, 2019.

[YK15] Fisher Yu and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." *arXiv:1511.07122*, 2015.

[YKS17] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. "Quicksilver: Fast predictive image registration–a deep learning approach." *NeuroImage*, **158**:378–396, 2017.

[YLJ19] Sheng Yang, Guosheng Lin, Qiuping Jiang, and Weisi Lin. "A dilated inception network for visual saliency prediction." *IEEE Trans. Multimedia*, 2019.

[YV18] Varduhi Yeghiazaryan and Irina D Voiculescu. "Family of boundary overlap metrics for the evaluation of medical image segmentation." *J. Med. Imaging*, **5**(1):015006, 2018.

[YXR18] Pingkun Yan, Sheng Xu, Ardeshir R Rastinehad, and Brad J Wood. "Adversarial image registration with application for MR and TRUS image fusion." In *Proc. Int. Workshop Mach. Learn. Med. Imag.*, pp. 197–204, 2018.

[ZDC19] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. "Recursive cascaded networks for unsupervised medical image registration." In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 10600–10610, 2019.

[Zha18] Jun Zhang. "Inverse-consistent deep networks for unsupervised deformable image registration." *arXiv:1809.03443*, 2018.

[ZHW19] You Zhang, Xiaokun Huang, and Jing Wang. "Advanced 4-dimensional cone-beam computed tomography reconstruction by combining motion estimation, motion-compensated reconstruction, biomechanical modeling and deep learning." *Vis. Comput. Ind. Biomed. Art*, **2**(1):1–15, 2019.

[ZHX21] Wentao Zhu, Yufang Huang, Daguang Xu, Zhen Qian, Wei Fan, and Xiaohui Xie. "Test-Time Training for Deformable Multi-Scale Image Registration." *arXiv preprint arXiv:2103.13578*, 2021.

[ZLL19] Shengyu Zhao, Tingfung Lau, Ji Luo, I Eric, Chao Chang, and Yan Xu. "Unsupervised 3D end-to-end medical image registration with volume tweening network." *IEEE J. Biomed. Health Inform.*, **24**(5):1394–1404, 2019.

[ZLZ20] Jinwei Zhang, Zhe Liu, Shun Zhang, Hang Zhang, Pascal Spincemaille, Thanh D Nguyen, Mert R Sabuncu, and Yi Wang. "Fidelity imposed network edit (FINE) for solving ill-posed image reconstruction." *Neuroimage*, **211**:116579, 2020.

[ZTZ17] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. "Scale-adaptive convolutions for scene parsing." In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2031–2039, 2017.

[ZWY20] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, et al. "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation." *IEEE Trans. Med. Imag.*, **39**(7):2531–2540, 2020.

[ZZL20] Mo Zhang, Jie Zhao, Xiang Li, Li Zhang, and Quanzheng Li. "ASCNet: Adaptive-Scale Convolutional Neural Networks for Multi-Scale Feature Learning." In *Proc. IEEE 17th Int. Symp. Biomed. Imaging*, pp. 144–148. IEEE, 2020.