Medical Image Registration Using Deep Neural Networks

Zachary Michael Cieman Baum

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Medical Physics and Biomedical Engineering
University College London

July 17, 2023
I, Zachary Michael Cieman Baum, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.
Abstract

Registration is a fundamental problem in medical image analysis wherein images are transformed spatially to align corresponding anatomical structures in each image. Recently, the development of learning-based methods, which exploit deep neural networks and can outperform classical iterative methods, has received considerable interest from the research community. This interest is due in part to the substantially reduced computational requirements that learning-based methods have during inference, which makes them particularly well-suited to real-time registration applications. Despite these successes, learning-based methods can perform poorly when applied to images from different modalities where intensity characteristics can vary greatly, such as in magnetic resonance and ultrasound imaging. Moreover, registration performance is often demonstrated on well-curated datasets, closely matching the distribution of the training data. This makes it difficult to determine whether demonstrated performance accurately represents the generalization and robustness required for clinical use.

This thesis presents learning-based methods which address the aforementioned difficulties by utilizing intuitive point-set-based representations, user interaction and meta-learning-based training strategies. Primarily, this is demonstrated with a focus on the non-rigid registration of 3D magnetic resonance imaging to sparse 2D transrectal ultrasound images to assist in the delivery of targeted prostate biopsies. While conventional systematic prostate biopsy methods can require many samples to be taken to confidently produce a diagnosis, tumor-targeted approaches have shown improved patient, diagnostic, and disease management outcomes with fewer samples. However, the available intraoperative transrectal ultrasound imaging alone
is insufficient for accurate targeted guidance. As such, this exemplar application is used to illustrate the effectiveness of sparse, interactively-acquired ultrasound imaging for real-time, interventional registration. The presented methods are found to improve registration accuracy, relative to state-of-the-art, with substantially lower computation time and require a fraction of the data at inference. As a result, these methods are particularly attractive given their potential for real-time registration in interventional applications.
Impact Statement

This thesis describes novel methods for medical image registration, a key enabling image analysis technique in healthcare settings. With the ever-increasing number of medical images being acquired and used for diagnosis, treatment planning, and surgical guidance; the development of fast, robust, and accurate registration methodologies has become increasingly important.

Within academia, this work addresses several notable challenges associated with learning-based multimodal and interactive registration techniques. Specifically, this thesis presents the first learning-based generalizable ‘model-free’ and non-rigid point-set registration method (Chapter 3). Additionally, the first adaptive, learning-based method for interactive registration is presented in Chapter 4 and further extended in Chapter 5. The desire with the introduction of such methods is that this thesis may not only further advance the existing state-of-the-art, but also aid in the development of other general-purpose registration methods. By contributing through multiple publications at the intersections of technical fields, such as deep learning and computer vision, and clinically-focused areas, such as prostate cancer, prostate biopsy guidance, and scoliosis quantification, this thesis has the potential to enable and create new avenues for research. Several interesting directions for such future work are discussed in Chapter 6.

Beyond academia, this work has implications for prostate cancer diagnosis: In current clinical practice, prostate biopsies are used to assist in the initial diagnosis and are additionally performed as a regular part of active surveillance programs. Though the sampling process is systematic, it relies greatly on the hand-eye coordination of the clinician, among other factors. However, if the prostate is undersampled,
clinically significant tumors can be missed and misdiagnosis can occur, leading to under-treatment of the patient. As a result, more samples are usually acquired, which can introduce complications for patients. The developments and evidence found through this research, whereby fast, data-efficient and adaptable methods for registration can deliver sufficiently accurate registration, are of critical importance in this instance. Migrating from current clinical processes and procedures where a systematic approach for prostate cancer biopsy is the standard, to an approach that is targeted and guided by fused preoperative and intraoperative imaging – without increasing the number of samples required or the length of the procedure – has the potential to make prostate biopsy procedures more effective and efficient. With nearly 100,000 prostate biopsies occurring in the United Kingdom each year by reducing the number of unnecessary biopsy samples; every minute saved during the procedure, and every complication, infection, or re-admission avoided by requiring fewer samples to be taken during the biopsy may result in considerable savings, financially and in working hours, to the health care system and help improve the quality of life of prostate cancer patients.
I must first thank and express sincere appreciation to my supervisors Dr. Dean Barratt and Dr. Yipeng Hu. Their guidance, and passion for the field are inspiring. It gives me great pleasure to have worked under and alongside them for the past several years. They are patient teachers and incredible mentors – always willing to explain, then re-explain, simple or complicated concepts over and over. Their knowledge of technical and clinical context, their commitment to me and my research, and their immense support has not gone unnoticed and will be ever-appreciated.

I would like to specifically thank Dr. Yipeng Hu for the innumerable supervision meetings, back-of-the-napkin math explanations, and responsiveness outside of regular working hours – for questions that were likely never as important as I thought. His enthusiasm and willingness to immerse me in things to do outside of my research afforded me countless opportunities to lecture and provide teaching assistance on some of my favorite topics, supervise students, run seminars series, organize international workshops, and take part in other research efforts and work which complimented and augmented my work. This thesis would not have been possible without his help, effort, guidance, mentorship, cheerleading and willingness to get his foot in the door for me when I didn’t think I could.

Other members of the Computer Assisted Navigation, Diagnosis, and Intervention (CANDI) group cannot go without my gratitude and appreciation, specifically; Dr. Matt Clarkson, Dr. Ester Bonmati, Dr. Alexander Grimwood, Dr. Zhe Min, Nina Montaña-Brown, Iani Gayo, Mark Pinnock, Shaheer Saeed, and Qianye Yang have all been incredible colleagues, co-authors, co-organizers, and friends throughout this journey. Working with them throughout my time at UCL has been an absolute
I’ve had the opportunity to work with so many other faculty, staff, and students within the Wellcome / EPSRC Centre for Interventional and Surgical Sciences, the Centre for Medical Image Computing, and the Department of Medical Physics and Biomedical Engineering at UCL – as well as many collaborators from outside of UCL. In particular, I’d like to thank Dr. Su-Lin Lee, Dr. Claudia Gandini Wheeler-Kingshott, Dr. Jamie McClelland, Dr. Tamas Ungi, and Dr. Christopher Schlenger, in addition to those in the CANDI group, for their thoughtful discussion, guidance, insight, and support in many research- and non-research-related endeavors.

I would like to acknowledge the funding sources which have supported, directly or indirectly, my PhD: the Natural Sciences and Engineering Research Council of Canada Postgraduate Scholarships-Doctoral Program, the University College London Overseas and Graduate Research Scholarships, the Wellcome Trust, the Engineering and Physical Sciences Research Council (EPSRC), and Cancer Research UK.

While not funded by the EPSRC Centre for Doctoral Training in Intelligent, Integrated Imaging in Healthcare, I would like to extend thanks for the involvement I was afforded in activities and events as an Aligned Student. It is through these events that I was able to meet many incredible, life-long friends and colleagues.

My friends ‘back home’ – Vinyas, Nuwan, Brandon, and Grace, my friends in London – Nina, Alkisti, Emma, and Ed, my family, and Sarah; without you, this never would have been finished. Your support and encouragement through all of the peaks and troughs have made the toughest parts of this PhD way less tough. Thank you.
UCL Research Paper Declaration Form: referencing the doctoral candidate’s own published work(s)

1. For a research manuscript that has already been published:

(a) What is the title of the manuscript?
   Meta-Learning Initializations for Interactive Medical Image Registration

(b) Please include a link to or doi for the work:
   https://doi.org/10.1109/TMI.2022.3218147

(c) Where was the work published?
   IEEE Transactions on Medical Imaging

(d) Who published the work?
   IEEE

(e) When was the work published?
   2/11/2022

(f) List the manuscript’s authors in the order they appear on the publication:
   Zachary M C Baum, Yipeng Hu, Dean C Barratt

(g) Was the work peer reviewed?
   Yes

(h) Have you retained the copyright?
   Yes

(i) Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)?
   https://arxiv.org/abs/2210.15371

2. For multi-authored work, please give a statement of contribution covering all authors:

Zachary M C Baum: Conceptualization, Methodology, Data curation, Software, Validation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing.
Acknowledgements

Yipeng Hu: Conceptualization, Data curation, Writing – review & editing, Supervision.

Dean C Barratt: Supervision.

3. In which chapter(s) of your thesis can this material be found?
   Chapter 4

e-Signatures confirming that the information above is accurate:

Candidate:
Date:

Supervisor/Senior Author signature:
Date:
UCL Research Paper Declaration Form: referencing the doctoral candidate’s own published work(s)

1. For a research manuscript that has already been published:

   (a) **What is the title of the manuscript?**
       Meta-Registration: Learning Test-Time Optimization for Single-Pair Image Registration

   (b) **Please include a link to or doi for the work:**
       https://doi.org/10.1007/978-3-031-16902-1_16

   (c) **Where was the work published?**
       ASMUS 2022: Simplifying Medical Ultrasound

   (d) **Who published the work?**
       Springer

   (e) **When was the work published?**
       9/15/2022

   (f) **List the manuscript’s authors in the order they appear on the publication:**
       Zachary M C Baum, Yipeng Hu, Dean C Barratt

   (g) **Was the work peer reviewed?**
       Yes

   (h) **Have you retained the copyright?**
       Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)?**
       https://arxiv.org/abs/2207.10996

2. For multi-authored work, please give a statement of contribution covering all authors:

   **Zachary M C Baum:** Conceptualization, Methodology, Data curation, Software, Validation, Formal analysis, Visualization, Writing – original draft,
Writing – review & editing.

Yipeng Hu: Conceptualization, Data curation, Writing – review & editing, Supervision.

Dean C Barratt: Supervision.

3. In which chapter(s) of your thesis can this material be found?
   Chapter 5

E-Signatures confirming that the information above is accurate:

Candidate:
Date:

Supervisor/Senior Author signature:
Date:
UCL Research Paper Declaration Form: referencing the doctoral candidate’s own published work(s)

1. For a research manuscript that has already been published:

   (a) **What is the title of the manuscript?**
       Learning Generalized Non-Rigid Multimodal Biomedical Image Registration from Generic Point Set Data

   (b) **Please include a link to or doi for the work:**
       https://doi.org/10.1007/978-3-031-16902-1_14

   (c) **Where was the work published?**
       ASMUS 2022: Simplifying Medical Ultrasound

   (d) **Who published the work?**
       Springer

   (e) **When was the work published?**
       9/15/2022

   (f) **List the manuscript’s authors in the order they appear on the publication:**
       Zachary M C Baum, Tamas Ungi, Christopher Schlenger, Yipeng Hu, Dean C Barratt

   (g) **Was the work peer reviewed?**
       Yes

   (h) **Have you retained the copyright?**
       Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)?**
       https://arxiv.org/abs/2207.10994

2. For multi-authored work, please give a statement of contribution covering all authors:

   Zachary M C Baum: Conceptualization, Methodology, Data curation, Soft-
ware, Validation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing.

Tamas Ungi: Data curation, Writing – review & editing.

Christopher Schlenger: Data curation, Writing – review & editing.

Yipeng Hu: Conceptualization, Writing – review & editing, Supervision.

Dean C Barratt: Writing – review & editing, Supervision.

3. In which chapter(s) of your thesis can this material be found?

Chapter 3

e-Signatures confirming that the information above is accurate:

Candidate:

Date:

Supervisor/Senior Author signature:

Date:
UCL Research Paper Declaration Form: referencing the doctoral candidate’s own published work(s)

1. For a research manuscript that has already been published:

(a) **What is the title of the manuscript?**
Real-Time Multimodal Image Registration with Partial Intraoperative Point-Set Data

(b) **Please include a link to or doi for the work:**
https://doi.org/10.1016/j.media.2021.102231

(c) **Where was the work published?**
Medical Image Analysis

(d) **Who published the work?**
Elsevier

(e) **When was the work published?**
1/12/2021

(f) **List the manuscript’s authors in the order they appear on the publication:**
Zachary M C Baum, Yipeng Hu, Dean C Barratt

(g) **Was the work peer reviewed?**
Yes

(h) **Have you retained the copyright?**
Yes

(i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)?**
https://arxiv.org/abs/2109.05023

2. For multi-authored work, please give a statement of contribution covering all authors:

**Zachary M C Baum:** Conceptualization, Methodology, Data curation, Software, Validation, Formal analysis, Visualization, Writing – original draft,
Acknowledgements

Writing – review & editing.

Yipeng Hu: Conceptualization, Data curation, Writing – review & editing, Supervision.

Dean C Barratt: Writing – review & editing, Supervision.

3. In which chapter(s) of your thesis can this material be found?
   Chapter 3

E-Signatures confirming that the information above is accurate:

Candidate:
Date:

Supervisor/Senior Author signature:
Date:
UCL Research Paper Declaration Form: referencing the doctoral candidate’s own published work(s)

1. For a research manuscript that has already been published:

   (a) **What is the title of the manuscript?**
       Multimodality Biomedical Image Registration Using Free Point Transformer Networks

   (b) **Please include a link to or doi for the work:**
       https://doi.org/10.1007/978-3-030-60334-2_12

   (c) **Where was the work published?**
       ASMUS 2020, PIPPI 2020: Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis

   (d) **Who published the work?**
       Springer

   (e) **When was the work published?**
       1/10/2020

   (f) **List the manuscript’s authors in the order they appear on the publication:**
       Zachary M C Baum, Yipeng Hu, Dean C Barratt

   (g) **Was the work peer reviewed?**
       Yes

   (h) **Have you retained the copyright?**
       Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)?**

2. For multi-authored work, please give a statement of contribution covering all authors:

   **Zachary M C Baum:** Conceptualization, Methodology, Data curation, Soft-
ware, Validation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing.

**Yipeng Hu**: Conceptualization, Data curation, Writing – review & editing, Supervision.

**Dean C Barratt**: Writing – review & editing, Supervision.

3. **In which chapter(s) of your thesis can this material be found?**
   
   Chapter 3

---

**e-Signatures confirming that the information above is accurate:**

**Candidate:**

**Date:**

**Supervisor/Senior Author signature:**

**Date:**
Contents

1 Background ........................................... 42
   1.1 Prostate Cancer .................................... 42
       1.1.1 Prostate Anatomy .............................. 43
       1.1.2 Diagnosis, Staging, Management and Treatment 44
       1.1.3 Prostate Imaging .............................. 53
   1.2 Medical Image Registration ......................... 63
       1.2.1 Overview .................................. 63
       1.2.2 Transformation Models .................... 64
       1.2.3 Similarity Measures ......................... 65
       1.2.4 Optimization Methods ...................... 67
   1.3 Deep Learning ..................................... 68
       1.3.1 Overview .................................. 68
       1.3.2 Datasets and Data Considerations for Deep Learning 73
       1.3.3 Deep Learning in Medical Imaging ............ 75
       1.3.4 Deep Learning in Medical Image Registration .. 76
   1.4 Medical Image Registration in Prostate Cancer Applications 79

2 Motivation and Objectives ............................ 82
   2.1 Clinical Motivations .............................. 82
   2.2 Thesis Objectives ................................ 84
   2.3 Thesis Structure ................................ 85
4.1 Introduction ................................................. 132
  4.1.1 Novel Training Paradigms for Learning-Based Medical Image Registration ............................................. 132
  4.1.2 Interactive Machine Learning ............................. 133
  4.1.3 Gradient-Based Meta-Learning ......................... 135
4.2 Contributions .............................................. 136
4.3 Interactive Medical Image Registration with Meta-Learning .... 137
  4.3.1 Learning-based Image Registration ..................... 138
  4.3.2 Interaction for Image Registration .................... 139
  4.3.3 Meta-Learning Interactive Initializations ............... 140
4.4 Methods ................................................ 141
  4.4.1 Images and Annotations as Interaction .................. 141
  4.4.2 Meta-Learning for Interactive Registration .......... 142
  4.4.3 Gradient-Based Meta-Learning Algorithms for Network Initialization .............................................. 144
  4.4.4 Exemplar Clinical Application ............................ 146
  4.4.5 Volume-to-Sparse Weakly Supervised Multimodal Image Registration .............................................. 147
  4.4.6 Interactive Acquisition of Labelled TRUS Images ........ 148
  4.4.7 Meta-Learning an Initialization with Reptile .......... 149
4.5 Experiments .............................................. 151
  4.5.1 Baseline Model Implementation and Training .............. 151
  4.5.2 Loss Functions ........................................ 152
  4.5.3 Data .................................................. 152
  4.5.4 Comparison with Meta-Learning Variants ................ 153
  4.5.5 Comparison with State-of-the-Art Approaches ............ 154
  4.5.6 Comparison with Non-Meta-Learning Approaches ........ 154
  4.5.7 Evaluation of Registration Methods .................... 155
4.6 Results ................................................ 156
  4.6.1 Baseline Performance .................................... 156
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6.2</td>
<td>Performance of Baseline Variants</td>
<td>158</td>
</tr>
<tr>
<td>4.6.3</td>
<td>Performance of State-of-the-Art Approaches</td>
<td>161</td>
</tr>
<tr>
<td>4.6.4</td>
<td>Performance of Non-Meta-Learning Approaches</td>
<td>163</td>
</tr>
<tr>
<td>4.7</td>
<td>Discussion</td>
<td>166</td>
</tr>
<tr>
<td>4.8</td>
<td>Conclusion</td>
<td>169</td>
</tr>
<tr>
<td>5</td>
<td>Test-Time Meta-Registration Optimization</td>
<td>170</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>170</td>
</tr>
<tr>
<td>5.2</td>
<td>Contributions</td>
<td>172</td>
</tr>
<tr>
<td>5.3</td>
<td>Methods</td>
<td>173</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Unsupervised Learning-Based Image Registration</td>
<td>173</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Test-Time Single-Pair Optimization</td>
<td>173</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Model-Agnostic Meta-Learned Test-Time Optimization</td>
<td>175</td>
</tr>
<tr>
<td>5.4</td>
<td>Experiments</td>
<td>176</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Meta-Registration Implementation</td>
<td>176</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Data</td>
<td>177</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Comparison Studies</td>
<td>177</td>
</tr>
<tr>
<td>5.4.4</td>
<td>Evaluation of Registration Methods</td>
<td>178</td>
</tr>
<tr>
<td>5.5</td>
<td>Results and Discussion</td>
<td>178</td>
</tr>
<tr>
<td>5.6</td>
<td>Conclusion</td>
<td>179</td>
</tr>
<tr>
<td>6</td>
<td>Conclusion</td>
<td>182</td>
</tr>
<tr>
<td>6.1</td>
<td>Future Work</td>
<td>183</td>
</tr>
<tr>
<td>6.1.1</td>
<td>Learning to Constrain ‘Free’ Transformations</td>
<td>184</td>
</tr>
<tr>
<td>6.1.2</td>
<td>Expanding Definitions of Interactivity</td>
<td>187</td>
</tr>
<tr>
<td>6.1.3</td>
<td>Enabling Open and Reproducible Science in Multimodal Medical Image Registration</td>
<td>190</td>
</tr>
</tbody>
</table>

Bibliography

193
List of Figures

1.1 Illustration of the prostate gland and surrounding anatomy. From Cancer Research UK, licensed under CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0), via Wikimedia Commons. 

1.2 Illustration of the prostate gland’s zonal structures and encompassed anatomy. From Mikael Häggström, M.D., licensed under CC0 1.0 (https://creativecommons.org/publicdomain/zero/1.0), via Wikimedia Commons. 

1.3 TRUS-guided prostate biopsy procedures. Subfigures (a) and (b) from Cancer Research UK, licensed under CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0), via Wikimedia Commons. 

1.4 The Gleason score system for grading prostate cancer. Lower scores indicate tightly packed glands, whereas spread-out glands indicate higher-grade cancers. From National Institutes of Health, Public domain, via Wikimedia Commons. 

1.5 TRUS-guided prostate brachytherapy procedures. Subfigures (a) and (b) from Cancer Research UK, licensed under CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0), via Wikimedia Commons. 

1.6 Different types of ultrasound transducers. 

1.7 Linear (left), curvilinear (middle), and radial (right) transducer scanning arrangements.
List of Figures

1.8 Sample sagitally- and axially-oriented TRUS images of the prostate gland acquired during a prostate biopsy procedure. 57

1.9 Sample sagitally- and axially-oriented MR images of the prostate gland and surrounding anatomical structures. From PROSTATE-MRI [54], licensed under CC BY 3.0 (https://creativecommons.org/licenses/by/3.0), via The Cancer Imaging Archive [55]. 58

1.10 T2-weighted imaging (left column), diffusion-weighted imaging (DWI) (middle column), and apparent diffusion coefficient (ADC) maps (right column) demonstrating prostate MR scans for three different patients. Yellow contours clinically significant indicate prostate cancer lesions. From [58], licensed under CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)/Cropped from original. 60

1.11 An ‘open bore’ MR scanner, where there is an opening on both sides of the subject. From United States Navy, Public domain, via Wikimedia Commons. 61

1.12 The flow of multiple inputs through a single perceptron unit, and associated activation function, into a single numerical output unit activation. 68

1.13 A MLP with four input nodes, two hidden layers comprised of eight nodes each, and a single numerical output. 69

1.14 A CNN with a three-layered MLP to predict a single numerical output. 72

3.1 Schematic representation of the FPT network design for non-rigid point-set registration. The global feature extraction module takes a target and source point-set and applies shared input and feature transformations to both, creating a global feature vector. The point transformation module serves as a per-point transformation of the source point-set by determining the displacement to be added in order to obtain the transformed point-set. 93
3.2 Rigid (top) and non-rigid (bottom) registrations with Free Point Transformer (FPT). Blue: source, yellow: target, green: transformed source. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104

3.3 Non-rigid registration with added Gaussian noise (a) and varying deformation (b) using FPT. Gaussian noise added at 0.01 SD (left), 0.02 SD (center), 0.04 SD (right). Deformation added at 0.1 SD (left), 0.2 SD (center), 0.4 SD (right). Blue: source, yellow: target, green: transformed source. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 105

3.4 Non-rigid ‘partial-to-full’ registrations at 25% occlusion of the source point-set (top) and ‘partial-to-partial’ registrations at 25% occlusion of the source and target point-sets (bottom) with FPT. Blue: source, yellow: target, green: transformed source. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 107

3.5 The generic spine atlas (left) and a sample ultrasound reconstruction (right) which were subsequently resampled into point-sets to be registered using FPT, wherein the generic spine becomes is used as the source point-set and the reconstruction is used as the target. Figure from [150]. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 109

3.6 Scatter plot of per-patient TxA measured from the deformed spine models vs. X-ray-based measurements. Figure from [150]. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110

3.7 Sample anterior-posterior (top) and lateral (bottom) visualization from patient data. Left to right: X-ray image, 3D reconstruction, deformed spine model, deformed spine model overlaid on 3D reconstruction, deformed spine model overlaid on the X-ray image. Solid lines show measurements in X-ray images, dashed lines show measurements on deformed spine models. Figure from [150]. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111

3.8 Point-sets depicting (from left to right) the anterior, right, posterior, and left views of a prostate volume obtained from a segmented TRUS (top) and MR volume (bottom) for one pair of patient images. Figure from [120]. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 113
3.9 Schematic representation of the FPT network design for non-rigid point-set registration as applied to the MR-TRUS fusion task. Figure from [120]. .............................................................. 114

3.10 Schematic representation of the G-RBF network design for non-rigid point-set registration. Similarly to FPT, the global feature extraction module takes a target and source point-set and applies shared input and feature transformations to both, creating a global feature vector. The G-RBF point transformation module applies a non-rigid transformation using the predicted control points and spline coefficients, as in Eq. 3.10, to obtain the transformed point-set. Figure from [120]. .............................................................. 116

3.11 Illustration of possible TRUS images acquired from the TRUS transducer. Images are captured in the sagittal plane (left) and are shown with other image slices that may be acquired in each of the three scenarios, from an axial view (right). Figure from [120]. ......... 118

3.12 Illustration of the Slice Centroid Distance (left) and Slice Span (right) in two simulated instances of the possible scenarios. The green point indicates the centroid of the points in the selected image slices, and the blue point indicates the centroid of the entire prostate volume. Figure from [120]. .............................................................. 119

3.13 Example image slices illustrating registration results from three different cases, shown in the transverse plane. Each image shows the original TRUS image with the transformed source (MR) contours (red) superimposed onto the target (TRUS) contours (green). Columns indicate registrations from each method; from left to right: FPT-Chamfer, FPT-GMM, G-RBF-Chamfer, G-RBF-GMM, ICP, and CPD. Figure from [120]. .............................................................. 122
3.14 MR-TRUS prostate glands showing the overlap of the transformed source (MR) and target (TRUS) point-sets. The green shape illustrates the target point-set, while the red shape illustrates the transformed source point-set. The first column shows the source and target after center-alignment. The remaining columns show registrations from various methods; from left to right: FPT-Chamfer, FPT-Chamfer trained with Scenario 1 image slices, FPT-Chamfer trained with Scenario 2 image slices, and FPT-Chamfer trained with Scenario 3 image slices. Figure from [120].

3.15 Example image slices illustrating registration results from three different cases, shown in the transverse plane. Each image shows the original TRUS image with the transformed source (MR) contours (red) superimposed onto the target (TRUS) contours (green). Columns indicate registrations from FPT-Chamfer for registrations with; from left to right; full volumes, Scenario 1, Scenario 2, and Scenario 3. Figure from [120].

3.16 Boxplots of the root-mean-square TRE per patient for the MR to TRUS and MR to partial TRUS registrations obtained from FPT-Chamfer, G-RBF-Chamfer, ICP, and CPD with 4,096 points. Figure from [120].

4.1 A schematic of conventional machine learning and interactive machine learning approaches at inference. In the conventional approach (top), previously unseen test data is passed through a trained model to produce a prediction. In contrast, the interactive machine learning approach (bottom) leverages user interactions to provide feedback to the model, based on the intermediate prediction, to obtain a more accurate prediction.
4.2 Diagram illustrating one episode of task-level learning wherein a task \((s_m^n, t_m^n)_n\) is sampled. For each interaction of the \(k\) interactions in the sampled task, the image pair \(\{x_{s_m^n}\}, \{x_{t_m^n}\}\) is coupled with an associated annotation \(\{\ell_{s_m^n}\}, \{\ell_{t_m^n}\}\). Figure from [207]. 145

4.3 Schematic representation of a framework for interactive medical image registration with meta-learning. A learning-based registration method is trained over multiple episodes during the meta-training phase (left) to learn an initialization for adaptation at inference. During each task-level learning episode, a task is sampled; comprising a set of image and annotation pairs. Following each episode, the meta-update (red arrow) updates the learning-based registration method based on the direction (dashed line) of the \(k\) task-level learning gradients (white arrows) and continues onward from the previously learned gradients (blue arrows). Once training is complete, the registration method is optimized in the meta-test phases (right). Here, few-shot learning on the target data, coupled with user-defined interactions, yields a fine-tuned registration model. Figure from [207]. 146
4.4 Schematic representation of the proposed framework for interactive medical image registration with meta-learning as applied to a weakly-supervised volume-to-sparse prostate MR-TRUS registration problem. The weakly-supervised learner is trained over multiple episodes during the meta-training phase (left) to learn an initialization for adaptation at inference. During each task-level learning episode, a task is sampled; comprising a set of images, labels, and some number of frames $F$. After each episode, the meta-update step updates the learner using the Reptile algorithm based on the task-level learning gradients. Once training is complete, the learner is optimized during the meta-test phases (right). Here, interactively-acquired, patient-specific, data is coupled with few-shot learning to yield a fine-tuned registration model, in real-time, as the TRUS image acquisition occurs. Figure from [207].

4.5 Illustration of possible TRUS images acquired from the TRUS transducer in the presented clinical scenario. Newly acquired images (dashed lines) are captured in the sagittal plane (left) and are shown with the previously acquired images (solid lines) throughout one continuous ‘sweep’ through the prostate with the TRUS probe until full coverage of the prostate has been obtained. Figure from [207].
4.6 Example image slice from one test case. The left-most column contains image slices from the source MR volume. The right-most column contains the corresponding target TRUS image slice. Other columns present the warped source MR image, resulting DDF, alternating vertical slices of the warped MR and target TRUS image, and warped MR prostate gland contour (Red) overlaid on the target TRUS prostate gland contour (Green), using the Baseline at a given shot of training during few-shot learning, with $F$ frames (e.g. the 2nd column indicates no few-shot learning, i.e. using the learned initialization with no fine-tuning, and two input frames, the 3rd column indicates one-shot of learning and three input frames, etc.). Figure from [207].

4.7 Tukey’s boxplots of TRE for the Baseline and all variants in the MR-TRUS registration experiment. Whiskers indicate 10th and 90th percentiles. Results are presented for registrations with 10 frames unless otherwise indicated. For $F_{\text{max}}$ variants, results are presented at 10 frames for direct comparison to the Baseline and other variants, and also with the number of frames corresponding to the value of $F_{\text{max}}$ used in training. Figure from [207].

4.8 Example image slices from one test case. The left-most column contains image slices from the source MR volume and the corresponding target TRUS image slice. Other columns present the warped source MR image, resulting DDF, alternating vertical slices of the warped MR and target TRUS image, and warped MR prostate gland contour (Red) overlaid on the target TRUS prostate gland contour (Green), using the above-labelled network; either the Baseline or one of its variants. Figure from [207].

4.9 Tukey’s boxplots of TRE for the Baseline and all state-of-the-art methods in the MR-TRUS registration experiment. Whiskers indicate 10th and 90th percentiles. Figure from [207].
4.10 Example image slices from two test cases, shown on the left and right, respectively. The left-most column contains image slices from the source MR volume and the corresponding target TRUS image slice. Other columns present the warped source MR image, resulting DDF, alternating vertical slices of the warped MR and target TRUS image, and warped MR prostate gland contour (Red) overlaid on the target TRUS prostate gland contour (Green), using the above-labelled network; either the Baseline or one of the state-of-the-art methods. Figure from [207]. 165

4.11 Tukey’s boxplots of TRE for the Baseline and all non-meta-learning methods in the MR-TRUS registration experiment. Whiskers indicate 10th and 90th percentiles. Results for the Baseline presented for registrations with 10 frames, and indicated explicitly for other methods, whereby the number of frames indicated is the number of frames used in training and evaluation. Figure from [207]. 166

4.12 Example image slices from one test case. The left-most column contains image slices from the source MR volume and the corresponding target TRUS image slice. Other columns present the warped source MR image, resulting DDF, alternating vertical slices of the warped MR and target TRUS image, and warped MR prostate gland contour (Red) overlaid on the target TRUS prostate gland contour (Green), using the above-labelled network; either the Baseline or one of the non-meta-learning-based methods. Figure from [207]. 167
5.1 Schematic representation of the proposed unsupervised meta-registration method for single-pair test-time optimization. A learning-based registration is trained over multiple episodes during the training phase (left). In each episode, a pair of images is sampled and repeatedly registered. Following each episode, the meta-update updates the registration model based on the learned gradients from the episode which was just completed. Once training is complete, the registration model may be optimized at test-time for a single pair of images (right) using few-shot learning to yield a registration model optimized for a specific pair of input images. Figure from [248].

5.2 Tukey’s boxplots of DSC for all methods. Whiskers indicate 10th and 90th percentiles. Figure from [248].

5.3 Tukey’s boxplots of TRE for all methods. Whiskers indicate 10th and 90th percentiles. Figure from [248].

5.4 Example image slices from one test case. The left-most column contains image slices from the source and target images. Other columns present the warped image, a checkerboard of the warped and target images, the warped prostate gland contour (Red) overlaid on the target prostate gland contour (Green), and the resulting DDF using the above-labelled method. Figure from [248].
List of Tables

3.1 Registration performance of FPT and other methods on unseen objects in the ModelNet40 dataset. The lowest mean value in each section is bolded. .................................................. 104

3.2 Non-rigid registration performance on unseen objects with added Gaussian noise. .................................................. 105

3.3 Non-rigid registration performance on unseen objects with varying deformation. .................................................. 106

3.4 Partial registration performance of FPT and other methods on unseen objects. .................................................. 106

3.5 Qualitative metrics which describe the three clinical scenarios. Values: Mean ± SD mm .................................................. 119

3.6 Dc, Dh, and TRE for each method used and at each point-set size in the first MR-TRUS registration experiment. The lowest mean value in each section is bolded. Significant differences with respect to FPT-Chamfer are denoted with an asterisk (*), based on two-tailed paired t-tests at \( \alpha = 0.05 \). Values: Mean ± SD mm. .................... 121

3.7 Time to compute a single registration at a given point-set size for FPT, G-RBF, ICP and CPD. Values: Mean ± SD s .................... 122

3.8 Dc, Dh, and TRE for each method used and at each grouped registration threshold. Values: Mean ± SD mm .................... 123
3.9 $D_C$, $D_H$, and TRE for each method used in the partial MR-TRUS registration experiment. All point-sets are of size 4,096. The lowest mean value in each section is bolded. Significant differences with respect to FPT-Chamfer are denoted with an asterisk (*), based on two-tailed paired t-tests at $\alpha = 0.05$. Values: Mean $\pm$ SD mm.

4.1 Summary TRE and DSC for the Baseline network at each step of Few-Shot Learning in the Meta-Test phase in the MR-TRUS registration experiment. Mean values are presented $\pm$ SD. TRE is given in mm.

4.2 Summary TRE and DSC for the $k = 1$ and $k = 100$ variants at each step of Few-Shot Learning in the Meta-Test phase in the MR-TRUS registration experiment. Mean values are presented $\pm$ SD. TRE is given in mm.

4.3 Summary TRE and DSC for the $\beta^{\text{meta}} = 0.25$ and $\beta^{\text{meta}} = 1.0$ variants at each step of Few-Shot Learning in the Meta-Test phase in the MR-TRUS registration experiment. Mean values are presented $\pm$ SD. TRE is given in mm.

4.4 Summary TRE and DSC for the $F_{\text{max}} = 5$ and $F_{\text{max}} = 15$ variants at each step of Few-Shot Learning in the Meta-Test phase in the MR-TRUS registration experiment. Mean values are presented $\pm$ SD. TRE is given in mm.

5.1 Summary of DSC, TRE and computation time for the Meta-Registration method as compared to the other methods evaluated. All values are presented as Mean. DSC and TRE are presented $\pm$ SD.
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D</td>
<td>One-dimensional</td>
</tr>
<tr>
<td>2D</td>
<td>Two-dimensional</td>
</tr>
<tr>
<td>3D</td>
<td>Three-dimensional</td>
</tr>
<tr>
<td>ADC</td>
<td>Apparent Diffusion Coefficient</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>AS</td>
<td>Active Surveillance</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>CPD</td>
<td>Coherent Point Drift</td>
</tr>
<tr>
<td>CT</td>
<td>Computerized Tomography</td>
</tr>
<tr>
<td>DC</td>
<td>Chamfer Distance</td>
</tr>
<tr>
<td>DH</td>
<td>Hausdorff Distance</td>
</tr>
<tr>
<td>DCE</td>
<td>Dynamic Contrast-enhanced</td>
</tr>
<tr>
<td>DDF</td>
<td>Dense Displacement Field</td>
</tr>
<tr>
<td>DSC</td>
<td>Dice Similarity Coefficient</td>
</tr>
<tr>
<td>DWI</td>
<td>Diffusion-weighted Imaging</td>
</tr>
<tr>
<td>FEM</td>
<td>Finite Element Model</td>
</tr>
<tr>
<td>FPT</td>
<td>Free Point Transformer</td>
</tr>
<tr>
<td>G-RBF</td>
<td>Gaussian Radial Basis Function</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphics Processing Unit</td>
</tr>
<tr>
<td>ICP</td>
<td>Iterative Closest Point</td>
</tr>
<tr>
<td>IML</td>
<td>Interactive Machine Learning</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>IoU</td>
<td>Intersection Over Union</td>
</tr>
<tr>
<td>MAML</td>
<td>Model-Agnostic Meta-Learning</td>
</tr>
<tr>
<td>MI</td>
<td>Mutual Information</td>
</tr>
<tr>
<td>MIND</td>
<td>Modality Independent Neighbourhood Descriptors</td>
</tr>
<tr>
<td>MLP</td>
<td>Multi-layer Perceptron</td>
</tr>
<tr>
<td>mp-MRI</td>
<td>Multiparametric Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>MR</td>
<td>Magnetic Resonance</td>
</tr>
<tr>
<td>PINN</td>
<td>Physics-informed Neural Network</td>
</tr>
<tr>
<td>PSA</td>
<td>Prostate-specific Antigen</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial Basis Function</td>
</tr>
<tr>
<td>ReLU</td>
<td>Rectified Linear Unit</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SGD</td>
<td>Stochastic Gradient Descent</td>
</tr>
<tr>
<td>SSD</td>
<td>Sum Of Squared Differences</td>
</tr>
<tr>
<td>TPSRPM</td>
<td>Thin-Plate Spline Robust Point Matching</td>
</tr>
<tr>
<td>TRE</td>
<td>Target Registration Error</td>
</tr>
<tr>
<td>TRUS</td>
<td>Transrectal Ultrasound</td>
</tr>
<tr>
<td>TxA</td>
<td>Transverse Process Angle</td>
</tr>
<tr>
<td>US</td>
<td>Ultrasound</td>
</tr>
<tr>
<td>X-ray</td>
<td>X-ray Imaging</td>
</tr>
</tbody>
</table>
Publications

The work presented in this dissertation was completed under the supervision of Dr. Dean C. Barratt and Dr. Yipeng Hu, and appears in the following publications:


In parallel to the completion of the research presented in this thesis, the author contributed via research or research-related efforts, preparation of publication, or supervision to the following works:


scoliosis visualization and measurement,” *IEEE Transactions on Biomedical Engineering* (2020).
Chapter 1

Background

This thesis is broadly focused on the development of novel methods for medical image registration using deep neural networks, with a particular emphasis on the registration of multimodal imaging for prostate cancer biopsy guidance and prostate cancer therapies. Notably, this thesis presents the application of these same proposed methods to other tasks, such as the quantification of spinal deformity for scoliosis monitoring and interpatient image registration to better provide references for comparing and analyzing individual patient images. The inclusion of these secondary applications demonstrates early-stage efforts toward extending the research beyond the core focuses given in each chapter. However, despite the differences in application, these works are fundamentally linked through their technical and clinical backgrounds. As such, the background material which is relevant to the entire thesis is presented in this chapter. In each further chapter, the distinct context and a literature review required for those specific technical methods and clinical applications are presented.

1.1 Prostate Cancer

Prostate cancer is the most common cancer affecting men in the United Kingdom with over 57,000 new cases each year [1, 2], with approximately 1 in 8 men being diagnosed in their lifetimes [1, 3]. It is also the second most common cause of cancer-related death in the United Kingdom, accounting for 14% of all cancer-related deaths in men [3, 4]. Globally, an estimated 1.4 million new prostate cancer cases – and
1.1. Prostate Cancer

nearly 400,000 deaths occurred in 2020 [5]. Prostate cancer is the leading cancer diagnosis in over 100 countries, and the leading cause of cancer-related death in nearly 50 countries [5]. The burden of prostate cancer is continually increasing given aging populations globally [6]. Over 95% of prostate cancers are adenocarcinomas, a type of cancer that forms in the tissue surrounding fluid-producing glands within the prostate [7].

1.1.1 Prostate Anatomy

The prostate is part of the male reproductive system found below the bladder and in front of the rectum (Figure 1.1) [8]. The prostate is formed of glandular and fibromuscular tissue, with the glandular tissue composing the interior of the gland, and the majority of the fibromuscular tissue comprising the outer capsule of the gland. The prostate has various functions, most importantly, the prostate produces seminal fluid – a component of semen. Furthermore, it has a role in hormone production – such as the conversion of testosterone to dihydrotestosterone – and helps to regulate the flow of urine given its location and that the urethra passes through the center of the prostate as it leaves the bladder and goes through the penis [8]. Prostate size changes with age, typically enlarging over time [8].

The prostate is split into three distinct anatomical zones; the peripheral zone, the central zone, and the transition zone (Figure 1.2). The peripheral zone, the largest zone – which comprises nearly 70% of the volume in most men – encompasses the outermost region of the prostate and encircles the central zone and most of the transition zone [9]. The peripheral zone is also where most prostate cancers are detected [9]. The central zone comprises approximately 25% of the prostate volume and surrounds the ejaculatory ducts. The transitional zone comprises approximately 5% of the prostate volume and is the most central part of the prostate [9]. The transitional zone encircles the distal end of the urethra. Additionally, the superior part of the prostate is called the base, while the lower, narrower part of the prostate is called the apex.
1.1. Prostate Cancer

1.1.2 Diagnosis, Staging, Management and Treatment

1.1.2.1 Prostate Cancer Diagnosis

Diagnostic tests can assist in finding early-stage prostate cancers through digital rectal examinations or testing for prostate-specific antigen (PSA) in a patient’s blood [7, 10]. In some regions, these are used as part of formal or informal screening programs [11]. During a digital rectal examination, a clinician will attempt to feel for any lumps or hardness which may indicate cancer. PSA is a protein that is produced by the prostate and by cancerous cells within the prostate. While a low level of PSA is common, increasing slightly with age as the prostate grows, a high PSA level can indicate prostate cancer or highly-treatable conditions such as prostatitis – the
inflammation of the prostate. However, such tests are not particularly accurate [12]. For example, a clinician may only be able to feel non-cancerous parts of the prostate, highlighting the subjective nature of digital rectal examinations. Furthermore, some men with high PSA levels may not have prostate cancer, and some men with low PSA levels may have prostate cancer. This can lead to false positives or false negatives, prompting unneeded interventions or giving a false sense of security to a patient who actually does have cancer.

As a result of a PSA test, examination, or other factors which suggest a patient may have prostate cancer, patients may receive diagnostic imaging, in the form of magnetic resonance (MR) imaging or multiparametric magnetic resonance imaging (mp-MRI) – the imaging and diagnostic methods for which are discussed in Section 1.1.3.2 – or a prostate biopsy to further their diagnosis. Under the National Institute for Health and Care Excellence (NICE) guidelines for prostate cancer diagnosis in the United Kingdom, mp-MRI – i.e. using multiple MR techniques in combination – should be offered to those with a suspected localized cancer [13]. A
follow-up prostate biopsy should be offered to those with intermediate- and high-risk cancers based on the MR results [13]. However, for patients with low-risk cancers, biopsy should only be offered after a discussion of the risks and benefits of the procedure [13]. This ruling out of the need for a prostate biopsy is important, despite the stress that lack of treatment may cause patients, given the low chance (approximately 10 – 30%) of clinically significant cancer being present despite low-risk imaging results, and the complications associated with prostate biopsy [13]. It is of note that the findings from MR are only indicative and do not constitute a definitive diagnosis in these cases, therefore a confirmatory biopsy is used to confirm more high-risk cancers where an MR indicates that the grade might be higher.

A prostate biopsy is a surgical procedure where at least 6, and often 10, small samples of prostate gland tissue are systematically removed from the left and right apex, base, and mid-gland [14]. These samples are subsequently examined to determine the differentiation between healthy and pathological tissue to map, diagnose, and grade the cancer. Prostate biopsies may be transrectal or transperineal: In transrectal biopsies, the needle is inserted through the rectal wall towards the prostate (Figure 1.3a). In transperineal biopsies, the needle is passed through the perineum and into the prostate (Figure 1.3b). The transperineal approach is considered more convenient due to its shorter learning curve, however, transrectal biopsies are more widely performed [15]. Transperineal biopsies permit access to the base of the prostate more easily; something which is often technically very difficult during transrectal biopsies. Furthermore, the transperineal approach is better suited to mapping biopsies where a relatively large number of samples are collected. Transrectal biopsies may be performed at a much lower cost and under local anesthetic at the patient’s bedside, unlike a transperineal biopsy which typically requires specialized equipment, staff, general anesthetic, and an operating room. However, transrectal biopsies require the biopsy needle to pass through the rectal wall, posing a risk of infection that is substantially greater than that of transperineal biopsies [15, 16]. In current practice, a specific type of ultrasound (US) imaging, namely transrectal ultrasound (TRUS) imaging, will always be used during the procedure to help guide
1.1. Prostate Cancer

the insertion of the biopsy needles towards known targets, in addition to traditional systematic sampling. Previously acquired MR imaging may also be used in conjunction with TRUS to perform targeted biopsies, having been shown to improve cancer detection rates, especially those with intermediate and higher risk cancers [17].

![Illustration of a TRUS-guided transrectal biopsy.](image1)

![Illustration of a TRUS-guided transperineal biopsy.](image2)

Figure 1.3: TRUS-guided prostate biopsy procedures. Subfigures (a) and (b) from Cancer Research UK, licensed under CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0), via Wikimedia Commons.

Notably, even when systematic biopsy sampling of the prostate is used during a biopsy, underdiagnosis may occur if the locations and trajectories of the needles do not accurately capture the full distribution of cancerous tissue [18]. This results in reported sensitivities of approximately 70–80% for most procedures [19], overall, potentially leading to false negative diagnoses in 3 in 10 patients. If there is justification for obtaining more samples, such as from a region where there is evidence of disease from pre-operative MR imaging, clinicians may acquire additional samples, with additional sampling improving sensitivity by as much as 30%, depending on which region of the prostate they are acquired from [20]. However, the collection of more samples increases the length of the procedure, may cause increased discomfort for the patient, increase the risk of infection, and increases the cost of histopathology [18, 21–23]. Worryingly, while between 1993 and 2011 as many as 4.2% and 0.8% of patients who underwent prostate biopsy experienced post-procedural fever.
and required hospital re-admission [22], these numbers are now as much as 17.5% and 6.3%, respectively [23], and are increasing in many developed countries, including the United Kingdom, Canada, and the United States [23]. It is likely that these numbers are increasing due to rising antimicrobial resistance [22–24], given that the increases were cumulative, appearing over time, and were not observed per patient between initial and repeat biopsies [23].

1.1.2.2 Prostate Cancer Grading and Staging

Cancer grading is used to describe the abnormality of the tumor cells in a given cancer, as compared to normal cells. One often used method for grading prostate cancer is the Gleason score [25, 26]: Gleason scores are determined by observing biopsy samples in the areas that make up the majority of the cancer in the sample, based on the amount of healthy tissue and the arrangement of any cancerous tissues [25]. While the numerical score attributed to each area may range from 1 to 5, scores 1 and 2 are rarely reported in practice as they indicate normal healthy tissue [26, 27]. The remaining scores indicate the amount of differentiation between the cancerous and healthy tissue, ranging from well-differentiated (3) to undifferentiated (5) and indicating low- and high-risk cancers, respectively [25, 26] (Figure 1.4).

To determine the Gleason score, a score is first assigned to the largest or most obviously cancerous region, and then to the other areas of growth as a whole. The sum of these two scores should add to a value between 6 and 10 to indicate the summed score, which is used to grade the cancer [25]. A score of 6 indicates well-differentiated cancer with cells that resemble healthy tissue. These cancers are likely to grow slowly and be lower risk. On the other extreme, a score of 10 indicates undifferentiated cancer with cells that look very different from healthy tissue. These cancers may grow quickly and aggressively and are high-risk.

Other grading systems, such as the International Society of Urological Pathology (ISUP) grade groups [28, 29], map Gleason scores onto a range of 1 to 5 instead of 2 to 10. Under the ISUP grade groups, the lowest grade is 1, not 6 – as in Gleason scoring. These changes stand to help patients better understand their diagnosis and disease progression, as well as their prognosis given the clear boundaries and
1.1. Prostate Cancer

Figure 1.4: The Gleason score system for grading prostate cancer. Lower scores indicate tightly packed glands, whereas spread-out glands indicate higher-grade cancers. From National Institutes of Health, Public domain, via Wikimedia Commons.

While the Gleason or ISUP score is an important aspect of grading prostate cancers, several other factors are commonly used alongside Gleason scoring to assist in determining the best treatment options. These factors include multiple diagnostic findings, such as PSA scores, examinations, findings in patient imaging, how many biopsy cores contained cancer, the concentration of cancer within those cores, and if the cancer has metastasized. Cancer staging also incorporates this kind of diagnostic information and is used to provide an overall indication of the size, growth rate, and anatomical spread of a given cancer. Staging for prostate cancer is often done using the TNM system [31], which is comprised of individual scores for the tumor (T), lymph nodes (N), and metastasis (M) (the spread of cancer beyond the organ in which it originated). The tumor can be classified into one of four stages (T1 – T4): T1 cancers are too small to be seen in imaging, and often cannot be felt through
1.1. Prostate Cancer

examination [31]; T2 and T3 cancers denote cancers that are contained within the prostate gland, and those which have only just broken through the prostate gland, respectively [31]; T4 cancers have spread beyond the prostate [31]. Stages T1 – T3 also have various sub-stages which more accurately describe the exact nature of the tumor size and spread within the prostate. The lymph nodes are classified as either N0 if there are no cancer cells in nearby lymph nodes, or N1 if there are cancer cells within the lymph nodes near the prostate [31]. Furthermore, two stages describe the metastasis of the cancer; M0 and M1. M0 indicates no spread beyond the prostate, whereas M1 indicates spread beyond the prostate. M1 has several sub-stages to indicate the location of the metastasis, such as into bone or other organs [31].

Different staging and risk classification protocols utilize the TNM system alongside other diagnostic information. In the United States, for example, the American Joint Committee on Cancer (AJCC) TNM system [32] is commonly used. In the United Kingdom and across Europe, the Cambridge Prognostic Group (CPG) system [33] is often employed. The AJCC TNM system incorporates the TNM system with PSA level and Gleason score to assign a stage between I and IV, where stages II through IV have several sub-stages with multiple criteria [32]. The CPG system also incorporates PSA and Gleason score to assign a stage between CPG 1 and CPG 5 [33]. There are no sub-stages within each stage in the CPG system. In stages I and II, and in stages CPG 1 through CPG 3, there may be no immediate treatment and Active Surveillance (AS) may be recommended. For the higher stages, or depending on patient age and general well-being, and the patient’s attitude towards treatment, localized or systemic treatments may be recommended.

1.1.2.3 Prostate Cancer Management and Treatment

AS is a treatment management option that is being increasingly recommended to patients diagnosed with low-risk disease, where their prostate cancer is likely to not be harmful during their lifetime. AS involves regular testing and monitoring, and there is the possibility that an intervention may occur should the disease progress. This permits a more managed and patient-specific approach. Several classification and scoring systems, such as the D’Amico Classification [34] or the Gleason Score
1.1. Prostate Cancer

[25, 26], present clinically significant thresholds and scoring processes for low-, intermediate-, and high-risk patients. Most patients who are recommended to receive AS are those for whom their cancer is not causing harm in the absence of treatment, and is localized to the prostate. Additionally, patients in these lower-risk categories who are selected for AS will have other low-risk indicators, such as a low PSA level. In short, those, and only those at high risk of a lower quality of life or reduced life expectancy would be recommended for treatment in practice unless the patient specifically requests an intervention (for instance, where the AS confers a significant mental health impact). This leads to a key advantage of AS, the need for potentially unnecessary surgery or other treatments and their accompanying side effects is removed. However, there are notable disadvantages as well. Mainly, the psychological stresses associated with a lack of treatment may be significant, and – in extreme cases – if there is significant disease progression in between examinations, the cancer may have spread beyond the prostate, preventing the effective use of less aggressive treatment options.

Throughout AS, patients will receive routine examinations, PSA tests, and prostate biopsies [7, 10]. For patients on AS, targeted interventions will only be offered if, or when, significant disease progression is observed, typically through changes to measures such as the patient’s Gleason Score. While AS provides most patients a positive outcome after delaying or indefinitely postponing interventions [35], many patients still undergo invasive or unnecessary treatments which can lead to significant side effects without yielding improved outcomes [36].

In cases where treatment is required due to the stage of cancer, or treatment is requested by the patient, the available treatment options for prostate cancer range from localized treatments, where a specific area of the body is targeted, to systemic treatments, where the entire body is targeted. For prostate cancer, local treatment options broadly include different types of surgery and radiation therapy [10]. Systematic treatments depend on the type, location, and stage of cancer, and may include therapies such as chemotherapy, hormone therapy, and immune therapy. Most early-stage cancers, especially those which have not spread into other parts of the body,
1.1. Prostate Cancer

Prostate cancers are often treated using localized, tissue-sparing treatment approaches.

Surgical options involve the partial or complete removal of the prostate, and often the surrounding lymph nodes, through a procedure called a prostatectomy. The removal of the lymph nodes, and their subsequent dissection, may be done to determine if the cancer has spread beyond the prostate and can help assess the severity of the cancer. Open, robotic, and laparoscopic prostatectomies are employed in practice to excise the prostate and lymph nodes. All approaches have been reported to cause similar side effects and complications, such as urinary incontinence or impotence [37–39]. Importantly, robotic and laparoscopic approaches have demonstrated shortened recovery times and cause less bleeding during surgery [38, 39].

Radiation therapy utilizes radiation to kill cancer cells, usually through a series of treatments over time. External-beam radiation therapy, stereotactic body radiation therapy, and brachytherapy are the more commonly utilized types of radiation therapy options for prostate cancer [10]. External-beam radiation therapy involves the use of focused beams of radiation directed at the cancer, delivered over multiple sessions [40]. Stereotactic body radiation therapy provides radiation from various angles while the patient is kept still through the use of a mesh-like shell so that the cancer receives a high dose of radiation, while the surrounding structures and tissues receive a much lower dose [41]. Brachytherapy involves the implantation of radioactive seeds directly into the cancer for less than an hour or for up to a year, in high-dose-rate (Figure 1.5a) and low-dose (Figure 1.5b) rate brachytherapy, respectively [42]. Both external-beam radiation therapy and brachytherapy may be offered independently for low-risk patients who opt for treatment in place of AS. For patients with intermediate- or high-risk cancers, external-beam radiation therapy and brachytherapy may be offered in tandem [43]. While the presence of common surgical side effects, such as urinary incontinence and impotence are lessened with radiation therapy [44], it is notable that for higher-risk cancers, prostatectomy has been shown to deliver more reliable patient outcomes [40]. Proton beam therapy, a newer form of radiotherapy, uses high-energy protons to kill the cancer cells [45]. While in some instances it reduces the damage to surrounding healthy tissues, the
outcomes are no better than more widely used radiation therapy methods, and so it is used less frequently than external-beam radiation to treat prostate cancer [46], however; prostate cancer treatment is a major application for proton beam therapy.  

![Illustration of a TRUS-guided high-dose brachytherapy procedure where radiation is delivered through wires inserted nearby and into the cancerous regions of the prostate for a short period of time.](Image 111x474 to 301x686)

![Illustration of a TRUS-guided low-dose brachytherapy procedure where radiation is delivered through implanted radioactive seeds which are inserted nearby and into the cancerous regions of the prostate over a long period of time.](Image 317x474 to 507x686)

**Figure 1.5:** TRUS-guided prostate brachytherapy procedures. Subfigures (a) and (b) from Cancer Research UK, licensed under CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0), via Wikimedia Commons.

### 1.1.3 Prostate Imaging

#### 1.1.3.1 Ultrasound Imaging

US is a widely-available non-ionizing imaging modality that produces cross-sectional two-dimensional (2D) or three-dimensional (3D) images. It is typically used for imaging internal structures, such as organs and other soft-tissue structures, but can also be used to image bones. US images are produced from the emission of high-frequency sound waves which are subsequently reflected and scattered at tissue interfaces where there is a change in acoustic impedance in the internal structure being imaged. Adopting a simplified pulse-echo model of ultrasound image formation, reflected sound waves, or echoes, detected by a linear-array, piezo-electric
ultrasound transducer, can be processed to form an image where each pixel in the
image corresponds to the location of the echo within the body, and the brightness of
each pixel corresponds to the ‘strength’ of the echo (i.e. amplitude of the received
signal) [47]. This type of US image is referred to as brightness mode, or ‘B-mode’,
and is one of the most commonly used US imaging methods [47]. In reality, small-
scale variations in the acoustic properties of tissue give rise to ultrasound energy
being scattered in all directions. Constructive and destructive interference of these
waves at the transducer surface gives rise to the characteristic ‘speckle’ that appears
in B-mode US images.

B-mode images are formed by placing an ultrasonic ‘transducer’, which acts
as both a source of and receiver of US waves, in contact with the subject’s skin.
The ultrasound waves are then emitted in short pulses, at a centre frequency, which
travel along a narrow beam into the body through the skin [47]. Most diagnostic US
frequencies range from 2 – 15 MHz. Lower frequencies are used for deeper structures,
and higher frequencies for more superficial ones because higher frequencies (with
shorter wavelengths) are more easily absorbed. Once inside the subject, the waves
are scattered and reflected off of different structures, which generates echoes. The
amount of resistance that the beams will encounter is called the acoustic impedance.
Acoustic impedance is a unique property of tissue, and it depends on the density of
the tissue and the velocity of the waves transmitted through the tissue. Relatively
large changes in acoustic impedance occur at tissue boundaries and are most easily
visually identified at the boundaries between tissues and between tissue and bone
because of the differences in densities. Because tissues are complex materials where
there is a continuous change in structure and density, the acoustic impedance varies
within a given tissue. This can give rise to distinct textures and patterns, even
within a single type of tissue. Once these echoes travel back to the transducer,
they are detected and used to generate the image. Based on the distance from the
transducer, calculated from the time taken to receive the echoes from the original
wave – assuming a constant speed of sound – and the strength of the echo when it
returns, the image can be reconstructed. As such, echoes from targets closest to the
source are received first, and echoes that describe structures at further depths return to the source after. This emission, reflection, and reconstruction process happens around 10 – 100 times per second in most modern US devices, giving a real-time depiction of the imaged area. The difference in acoustic impedance across tissue interfaces and within a given tissue is what is ultimately shown in a US image.

The relative brightness within a given tissue or structure in a US image is termed echogenicity. A tissue with higher echogenicity is ‘hyperechoic’, a tissue with lower echogenicity is ‘hypoechoic’, whereas tissues that have similar echogenicity are termed ‘isoechoic’. While many cancers are isoechoic or hypoechoic [48–50], hyperechoic cancers are also possible, especially in prostate cancer [49]. Though they are not as common, they are often more severe and can indicate a later stage of disease [49]. In the US imaging of 200 patients, Spajić et al. found hyperechoic lesions in 9.5% of patients, and 7.6% of the identified lesions were hyperechoic [49]. Additionally, while hypoechoic and isoechoic lesions had mean Gleason scores of 5.6 and 5.4, respectively, hyperechoic lesions had a mean Gleason score of 7.0 [49], suggesting that hyperechoic lesions are more likely to be cancerous. It is also notable that calcifications within the prostate, more common in older patients, are also hyperechoic, and may resemble cancers in US imaging.

While most B-mode transducers are designed for external use (Figure 1.6a), there exist several types of transducers which are made for intracavity use (Figure 1.6b). TRUS, commonly used for targeted prostate biopsies, as discussed in Section 1.1.2.1, often uses a radial arrangement of transducer elements, unlike the linear or curvilinear arrangements common in external transducers (Figure 1.7). While linear-array transducers give a consistent field of view throughout the entire image (Figure 1.8a), the radial arrangement is particularly suited to internal imaging as it gives a very wide field of view of up to 360° (Figure 1.8b), either by virtue of a circular transducer or through the rotation about the central axis of a smaller transducer [47].

TRUS is commonly used for prostate imaging given the prostate gland’s position anterior to the rectum, permitting clear visualization and evaluation of the prostate.
1.1. Prostate Cancer

(a) A linearly-arranged ultrasound transducer, typically used for externally imaging through direct contact with the subject’s skin (Ultrasonix L14-5/38 Linear array Ultrasound Transducer Probe, Ultrasonix Medical Corporation, Richmond, BC, Canada).

(b) A radially-arranged transrectal ultrasound transducer, used for internally imaging within the rectum (BK 8848 Endocavity Biplane Transducer Probe, BK Medical - a GE HealthCare Company, Burlington, MA, USA).

Figure 1.6: Different types of ultrasound transducers.

Figure 1.7: Linear (left), curvilinear (middle), and radial (right) transducer scanning arrangements.

in real-time (Figure 1.8) [51, 52]. However, most prostate cancers sampled during TRUS-guided biopsies are not easily visible in TRUS due to their isoechoic nature relative to other tissues in the prostate gland [49, 52]. Because of the difficulties associated with visualizing some prostate cancer pathologies in US and TRUS imaging, MR imaging is being increasingly used to assist in the detection and diagnosis of prostate cancers.

1.1.3.2 Magnetic Resonance Imaging

MR imaging is a 3D volumetric imaging modality that produces detailed anatomical images by exciting – and subsequently detecting the changes made to – protons
found in the water within tissues [53]. Through the use of powerful magnets, a strong magnetic field is produced which aligns protons in the subject’s tissues with that magnetic field. During image acquisition, a radiofrequency current is repeatedly pulsed through the subject. This causes the protons in tissues to excite, and start to spin out of equilibrium, causing a pull against the magnetic field. Between these pulses, the energy released through ‘relaxation’, as the protons realign with the magnetic field, is detected and measured. The relaxation time and energy release which are required to realign with the field are what give different tissues, structures, and organs their distinct appearances in MR imaging and this is measured by the scanner [53]. MR imaging is particularly well suited to soft tissues, however, it typically provides poor visualization of bony regions since there is little to no water within such structures unless specialized imaging techniques are used.

Precise differentiation between certain tissues can be obtained by modifying the timing of the sequences of radiofrequency pulses. The two most common MR sequences are called T1- and T2-weighted images. In general, T1-weighted images use shorter times between pulses, and T2-weighted images use longer times between pulses [53]. In practice, this means that T1-weighted images are more effective for
identifying fatty tissues and obtaining morphological information. For T2-weighted images, the visualization of inflammation and lesions (due to high water content), as well as the ability to assess the zonal anatomy of the prostate, is improved (Figure 1.9).

Figure 1.9: Sample sagittally- and axially-oriented MR images of the prostate gland and surrounding anatomical structures. From PROSTATE-MRI [54], licensed under CC BY 3.0 (https://creativecommons.org/licenses/by/3.0), via The Cancer Imaging Archive [55].

The morphological assessments possible with T2-weighted images can be combined with other MR imaging techniques through mp-MRI. Often, techniques such as diffusion-weighted imaging (DWI) and dynamic contrast-enhanced (DCE) imaging will be included as part of the diagnostic imaging process for prostate cancer [56]. A type of functional imaging, DWI exploits the variations in the motion of water by measuring the random Brownian motion of water molecules in tissue [53]. The contrast in DWI comes from the variation of motion in water, and the signal is inversely correlated to the degree of random motion in the water molecules. A magnetic field gradient is applied, dephasing the spin of the water molecules, and causing the spin phases to vary along the gradient direction. A second gradient is then applied in the opposite direction, to cancel the previous dephasing. This results in the moving water molecules experiencing a different gradient intensity,
while the non-moving water molecules experience the same gradient intensity. This leaves the moving water molecules dephased and the non-moving water molecules in phase. The higher motion then has a higher signal loss, and the lower motion has a lower signal loss. In most of the normal, glandular tissue throughout the prostate, water molecules move freely. Conversely, prostate cancer growth depletes normal tissue, replacing it with a more dense tissue, which restricts the movement of water molecules. These regions appear brighter in the acquired images due to the lower signal loss (Figure 1.10). While the combination of DWI and T2-weighted imaging improves diagnostic capabilities [57], DWI has relatively poor spatial resolution and specificity [57].

To overcome the resolution and specificity issues associated with DWI, quantitative maps may be extracted from multiple DWI with different weightings to form an apparent diffusion coefficient (ADC) map. The ADC map represents a measure of the magnitude of diffusion of water molecules within tissue at a given voxel location. An ADC map is obtained by acquiring several conventional DWI images with different weightings to obtain a “pure” diffusion coefficient at each voxel. While a minimum of two acquisitions are required to create an ADC map, the accuracy of the map can be improved by obtaining multiple acquisitions at different attenuation coefficients. In the ADC map, tissues appear with inverse intensity characteristics to DWI, giving normal tissue a brighter appearance and prostate cancer a darker appearance (given the smaller magnitude of diffusion) in the generated map (Figure 1.10).

DCE imaging utilizes T1-weighted imaging, acquired before, during, and after the administration of a contrast agent. The contrast most often used is gadolinium-based, which is administered intravenously. Due to the contrast agent’s ability to enhance the visibility of vasculature, and the highly-vasculature nature of tumors, DCE can be a useful tool for obtaining diagnostic information in certain circumstances (e.g. for risk stratification in the transition zone for ambiguous lesions [59]), but consideration must be given to whether the potential diagnostic benefits outweigh the risks of administering a contrast agent [59]. The blood vessels in tumors tend to be more porous and ‘open’ relative to normal tissue, which results in rapid and early
enhancement, as well as the early and rapid outflow of the contrast agent. In DCE imaging, gadolinium-rich regions reduce the relaxation time of the voxels, leading to a brighter image at a given time-point, which can indicate the presence of a tumor. However, given the toxicity of gadolinium, the need for additional scanning time, and the lack of significant value for diagnosis, the effectiveness and continued use of DCE imaging for prostate cancer diagnosis are debated [59].

Overall, while T2-weighted imaging on its own has diagnostic use, the combi-
nation of multiple mp-MRI sequences, such as those described above, can provide additional diagnostic value [56]. The use of T2-weighted imaging with DWI and the use of T2-weighted imaging with DCE imaging have both been shown to assist in more accurately detecting prostate cancer [60, 61]. Tanimoto et al. demonstrated a significant change in diagnostic accuracy for prostate cancer over analysis of 83 series of patient imaging when using T2-weighted imaging and DWI compared to solely T2-weighted imaging, where accuracies were 84% and 64%, respectively [61].

While US imaging requires direct contact with the subject’s skin, MR images may be obtained by placing the subject inside the bore of an MR scanner (Figure 1.11), which contains a large, superconducting electromagnet. Because MR image acquisition can take from seconds to many minutes, scanning sessions can take an hour or more, and subjects must remain still to prevent any motion artifacts in the images.

![Figure 1.11: An ‘open bore’ MR scanner, where there is an opening on both sides of the subject. From United States Navy, Public domain, via Wikimedia Commons.](image)

MR imaging has various roles within prostate cancer diagnosis and management, owing to its accurate imaging capabilities. One such use, MR-guided biopsy, offers
the ability to precisely target specific regions during diagnosis, as discussed in Section 1.1.2.1. However, MR also plays an invaluable role in cancer grading and staging, treatment planning, and management [62]: With respect to grading, in addition to PSA values and Gleason scores, MR assessment is an important part of risk classification systems [34] and provides accurate visualization of regional metastatic spread [62]. A widely used reporting system is the Prostate Imaging-Reporting and Data System (PI-RADS). PI-RADS is used for the evaluation of suspected prostate cancer based on findings from mp-MRI and is used to predict the chance that a given cancer is clinically significant [63]. Here, clinically significant means that the patient will likely require treatment and is defined as a Gleason score of at least 7, a tumor volume of over 0.5 mL, or any tumor growth which has progressed beyond the outside of the prostate [63, 64]. Under PI-RADS, lesions are given a score from 1 to 5, where 1 indicates a very low and 5 indicates a very high probability of clinically significant cancer being present. Lesions can be scored on T2-weighted imaging as well as in DWI or DCE. The scoring from each of these types of imaging also depends on whether the lesion is located in the transition- or peripheral zone of the prostate [63, 64].

MR imaging plays an increasingly important role in surgical and treatment planning, its use often being crucial for determining potential surgical candidates [65]. As tumors that are solely contained within the prostate meet the eligibility criteria for prostatectomy, MR-based assessment can be useful for determining suitability for the procedure [62]. Furthermore, MR has been used for patient selection for brachytherapy, as well as determining optimal planning for the placement of the radioactive seeds to appropriately target the tumor site within the prostate while avoiding critical structures such as the rectum and urethra [66]. During post-treatment disease management, MR has found uses for detecting tumor recurrence after prostatectomy [67, 68] and to assess patient and tumor response to different radiation therapies [62].
1.2 Medical Image Registration

1.2.1 Overview

Medical imaging technologies, such as US and MR, have become commonplace throughout many aspects of healthcare. They have fundamentally changed how diagnosis, treatment planning, and surgical guidance are performed, and continue to do so. Often in these types of clinical tasks, multiple images may be acquired. Importantly, these images are acquired at different times – minutes, days, or even months apart – using different modalities, and may even be from different patients (for population-based assessments). The need to simultaneously visualize and compare these images is apparent and is accomplished through a process called image registration; now a task of fundamental importance to the field. Medical image registration describes the process in which correspondence is established between a pair of medical images such that they may be spatially transformed to align the images themselves and the anatomical structures within each of the images [69].

Medical image registration methods are historically divided into intensity-based methods and feature-based methods [69, 70]: In the literature, intensity-based methods are distinguished according to whether the registration seeks to align images through the inherent patterns in image intensities (or image intensity distributions) through one or more similarity measures. This is done without the explicit use of extracted image features. Feature-based methods, on the other hand, are distinguished by whether the registration seeks to align images through the establishment of correspondence between features derived or extracted from the images, such as points, contours, or surfaces. These features may be extracted explicitly, for instance, through the manual or algorithm-based identification of organ boundaries and other anatomical landmarks.

An often-used case of medical image registration is so-called ‘pairwise’ registration, where two images are aligned. Here, one image is referred to as the ‘source’ image and the other is the ‘target’ image. The source image is spatially transformed to align with the target image through a process that requires a transformation model, a similarity metric, and an optimization method.
1.2. Medical Image Registration

Formally, a registration task may be described as the problem of determining the transformation $T$, which optimally aligns the corresponding pixels, voxels, or other features in image $x_{\text{source}}$, to those in image $x_{\text{target}}$ image and can be expressed as:

$$T(x_{\text{source}}) \approx x_{\text{target}} \quad (1.1)$$

We may define a similarity measure, discussed further in Subsection 1.2.3, in terms of $T$, $x_{\text{source}}$, and $x_{\text{target}}$ as a function $f_{\text{sim}}$ such that:

$$S = f_{\text{sim}}(T(x_{\text{source}}), x_{\text{target}}) \quad (1.2)$$

We may find the optimal parameters $\theta_{\text{optimal}}$ for a parametric transformation $T(\theta_{\text{optimal}})$ that maximizes the similarity measure of $f_{\text{sim}}$ by some optimization process – as described in Subsection 1.2.4:

$$\theta_{\text{optimal}} = \arg \max_{\theta_{\text{optimal}}} f_{\text{sim}}(T(x_{\text{source}}), x_{\text{target}}) \quad (1.3)$$

Parametric transformations are explored further, along with non-parametric transformations, in Subsection 1.2.2.

1.2.2 Transformation Models

Medical image registration algorithms may be differentiated by the transformation models they use to describe the spatial transformation which relates the source and target image spaces. Broadly, transformations can be described as either rigid or non-rigid. Rigid transformations are typically composed of rigid-body translations and rotations [69]. Affine transformations also include scaling and shearing [69]. Rigid transformations are inherently global and are applied to the entire image. As such, they are typically applied for the registration of rigid structures, or as part of a coarse, initial registration, ahead of an additional refinement step. These qualities make it difficult to model local differences between images, and the anatomical structures within them. In such instances, additional deformation is required, and this is often
defined or applied by using non-rigid transformations [71]. Non-rigid transformations can either be parametric or non-parametric, where parametric transformations are described by a model with a given number of parameters or functions, and non-parametric transformations allow each element of an image to be displaced independently. Thin-plate splines [72], B-splines [73, 74], and diffeomorphisms [74, 75] are examples of parametric non-rigid transformations, whereas elastic- [76, 77] and optical flow- or fluid-based [78] transformations are considered to be non-parametric.

1.2.3 Similarity Measures

Similarity measures are used to measure the extent of spatial alignment following a registration, typically via the assessment of two separate terms, combined into one similarity measure, also called a ‘cost function’. The first term measures some aspect of the relationship between voxel intensities for intensity-based image registration methods and some form of structural similarity for feature-based image registration methods. The second term is used to impose constraints or a form of regularization on the deformation which defines the registration. Numerous similarity and regularization measures exist for both intensity- and feature-based methods. Several common measures for each are described below.

1.2.3.1 Intensity-Based Similarity Measures

Many intensity-based similarities have been used in practice, however, most methods typically utilize the intensity differences or cross-correlations which may exist between the source and target images. Measures based on the intensity differences between the warped source and target images may be derived from the sum of squared differences (SSD) [79, 80], typically when the intensities are correlated, as in unimodal registration. The SSD aims to demonstrate similarity through the premise that when well registered, each voxel or pixel in the images should have the same intensity, giving a lower SSD. The use of cross-correlation operates in much the same way [81, 82], though instead of assuming identical intensities, the assumption is that there is a linear relationship between the intensities of similar structures in each image. Conversely to the SSD, most correlation-based measures
desire a higher value to indicate a better correlation between the images and in-turn a better registration.

While these measures are both often used in image registration, these tend to be more effective in unimodal registration tasks - where the source and target images share the same intensity distribution, such as MR-MR registration tasks. However, this is not always the case, possibly due to the noise which may be present in the image. One such example is in modalities such as US, where these measures can be less effective. Furthermore, in scenarios where multimodal image registration is required, e.g. MR-US, these measures may fail when computed on the images directly. An often-used method for multimodal (and unimodal) image registration is Mutual Information (MI), an information-theoretic measure that describes the mutual dependence between two given images [83–86]. In essence, MI describes how well each image can explain the other, where we assume there exists some deterministic function between each image, such that all information in each image is shared; a high value for the MI similarity measure should indicate a better registration between two images.

1.2.3.2 Feature-Based Similarity Measures

Feature-based similarities often involve minimizing distances between the source and target data through features, such as point-sets, contours, or binary segmentation maps, extracted from, or defined by the source and target images. Typically, these features will represent the region of interest (e.g. the organ). As with intensity-based registrations, the SSD may be used here, though it will define a spatial distance between corresponding points or structures, rather than the difference in intensities. Other metrics for spatial distances between image features, such as the Euclidean distance or Chamfer Distance ($D_C$), have been used to describe aggregate point-closeness [87, 88]. For both measures, a smaller distance typically indicates a better registration (i.e. a closer ‘fit’ between the source and target features). In the case of segmentation maps, overlap measures typically used for evaluating a given segmentation, such as the Dice similarity coefficient (DSC) or Intersection over Union (IoU), may be used to measure similarity in feature-based image registration
1.2. Medical Image Registration

processes. In such instances, the similarity or degree of overlap of the segmentation maps is determined, respectively. In both cases, a higher value for such overlap-based metrics indicates higher similarity, and likely a more accurate registration.

1.2.3.3 Regularization Measures

Multiple regularization measures exist for preserving smoothness or constraining the deformations in the registration. Existing approaches have used measures such as the L1 Norm of the displacement gradient [73], or the sum of squared first-order derivatives of the transformation [89]. However, the most common approaches utilize the second-order derivatives of the transformation to regularize the bending energy [73] or the Jacobian determinant to apply an incompressibility constraint to the registration [90].

1.2.4 Optimization Methods

On the assumption that the optimal value of a given similarity measure corresponds to correctly registered images, the goal of the optimization algorithm is to find, often iteratively, the transformation that maximizes or minimizes (depending on the measure) the similarity measure. While several optimization methods exist, methods and solutions which are based on gradient descent (or ascent) [73, 91, 92] and least-squares approaches [93, 94], such as the Procrustes method [69], are among the most commonly employed. Gradient descent is commonly applied in both intensity- and feature-based methods. In gradient descent, following the computation of the similarity and gradient, a step is taken along the search space in the direction of the computed gradient. This occurs iteratively until convergence is achieved. Least-squares methods are conventionally applied in feature-based methods. In least-squares approaches, the rigid transformation which minimizes distances between points is computed by determining the rotation and translation which define the optimal alignment.
1.3 Deep Learning

1.3.1 Overview

Deep learning represents a class of machine learning algorithms that are composed of processing layers that can learn high-level representations and features from data at various levels of abstraction [95]. Deep learning has enabled great improvements and redefined the state-of-the-art in numerous fields through its ability to universally approximate functions that relate complex datasets. Most deep learning methods are based on the Artificial Neural Network (ANN), a system built from perceptrons, the fundamental building block of ANNs. These perceptrons loosely model the connections in a brain to learn to perform a specific task without utilizing hand-engineered rules or heuristics and can be described mathematically as:

$$ a = \phi \left( \sum_{i=1}^{n} w_i x_i + b \right) = \phi \left( w^T x + b \right) $$

where $a$ is the unit’s activation, the layer $w$ is a vector of weights, $x$ is a vector of inputs, $b$ is the bias, and $\phi$ is the activation function. The flow of the inputs through a single perceptron unit is illustrated in Figure 1.12.

![Figure 1.12: The flow of multiple inputs through a single perceptron unit, and associated activation function, into a single numerical output unit activation.](image)

Activation functions map the weighted inputs of the perceptron to the unit’s output. Often, these functions are nonlinear and are designed to mimic the firing of biological neurons, such as the logistic sigmoid function: $\phi(x) = (1 + e^{-x})^{-1}$. 
1.3. Deep Learning

Given its smooth shape and range of \([0, 1]\), the logistic sigmoid function was more commonly used as an activation function before the introduction of rectifier-based functions, such as the Rectified Linear Unit (ReLU) [96] – computed as \(\varphi(x) = \max(x, 0)\). ReLU provides many benefits in deeper network architectures, such as sparse activations, better gradient propagation and efficient computation [96].

On its own, an individual perceptron has limited mapping ability. However, perceptrons can be assembled into networks comprising multiple layers, each with multiple perceptron units, referred to as a multi-layer perceptron (MLP), where the input is propagated layer-by-layer through the network (Figure 1.13). The first layer will receive input data, and the final layer will provide the inferred output. The layers between the input and output are referred to as hidden layers. Each given node in a layer is typically ‘fully connected’ in the sense that each node is connected to every node in the subsequent layer and permits the learning combinations of non-linear features within the neural network. This means that each perceptron unit has its own bias and weights for every pair of units in consecutive layers.

![Figure 1.13: A MLP with four input nodes, two hidden layers comprised of eight nodes each, and a single numerical output.](image)

Once assembled, these networks are capable of learning through a training process. This occurs in an iterative, epoch-based manner over many different training
1.3. Deep Learning

examples, where the number of epochs indicates the number of passes through the training data that will occur during training. Formally, this may be considered a type of minimization problem, where, given a neural network $f$ and learnable parameters $\theta$, the goal of a deep learning approach is to determine the optimal parameters of $\theta$, $\theta^*$ which minimizes an expected loss $\mathcal{L}_{\text{expected}}$ on data $(x,y)$, distributed with respect to the joint distribution of inputs and target labels $P(X,Y)$, as:

$$
\theta^* = \arg\min_{\theta} \mathcal{L}_{\text{expected}} = \arg\min_{\theta} \mathbb{E}_{(x,y) \sim P(X,Y)} \mathcal{L}(y, f(x; \theta))
$$

where $x \in X$ is an input and $y \in Y$ a corresponding target label. Notably, the learnable parameters $\theta$ may contain weights $w$ and biases $b$, as described above for a MLP.

It is also of note that the true distribution of $P(X,Y)$ is unknown. This means that we must alternatively find $\theta$ which empirically minimizes $\mathcal{L}_{\text{empirical}}$ and gives an approximate solution to $\theta^*$ as $\hat{\theta}$ on data $(x,y)$, distributed with respect to an available dataset $D$:

$$
\hat{\theta} = \arg\min_{\theta} \mathcal{L}_{\text{empirical}} = \arg\min_{\theta} \mathbb{E}_{(x,y) \sim D} \mathcal{L}(y, f(x; \theta))
$$

This learning objective is employed across several fundamental learning paradigms, of which supervised learning and unsupervised learning are among the most commonly employed. Both are described in the following sections. In most cases, supervised learning is used; however, a key requirement is so-called ‘labeled’ data, where the target output is known for each training input and defines the label for that input datum. Unsupervised learning, on the other hand, learns without any labeled data.

1.3.1.1 Supervised Learning

In supervised learning, the training is guided by a loss function, which determines the goodness-of-fit (e.g. the difference) between the network’s output (typically, a prediction generated by the neural network) and the target output for the given input data. The weights between each node in the network are then adjusted based on a learning rule and the computed value of the loss function. Over several training
epochs, the network’s output should converge to become more and more similar to the target output for the input data. Training may be terminated based on several criteria, such as a fixed number of training epochs or through achieving a pre-defined performance threshold on an external set of independent validation data. Supervised learning has the advantage of being able to learn how to explicitly model the data given to it and be able to give a specific prediction on new, unseen data, often generalizing well to unseen data which is within or near its training data distribution. However, supervised learning can be restricted in its knowledge and its ability to robustly infer anything about inputs that are greatly outside of its training data distribution. Additionally, supervised learning requires a large amount of labeled data to be effective, and the data must be labeled or reviewed by a human expert. This can be a time-consuming and expensive process and may be difficult to obtain in some cases, such as in the case of image segmentation where each pixel must be labeled.

1.3.1.2 Unsupervised Learning

In unsupervised learning, training is also guided by a loss function. However, instead of comparing the goodness-of-fit of the network’s output to the target output, the network’s output is typically compared to the input data. Like supervised learning, the weights between each node in the network are then adjusted based on a learning rule and the computed value of the loss function to, over several training epochs, improve the network’s ability to reconstruct the input data. Unlike supervised learning, unsupervised learning may be considered to be a type of discovery or organization task. Often, the network attempts to uncover patterns and representations in the data by learning to mimic or reconstruct its input. Notably, different networks may infer different patterns from the same data. Unsupervised learning is advantageous in many ways, two of the most salient being that it permits the learning of trends and patterns within the underlying data which humans may not have been able to infer, and the data does not require any labeling. Unfortunately, while these trends and patterns may be useful, they often require domain knowledge and expertise to fully utilize and understand. Additionally, unsupervised methods may be considered
more difficult to validate, as there is not necessarily a metric that corresponds to the effectiveness of a prediction in practice. Unsupervised learning is also more difficult to apply, as it requires a large amount of data to be effective.

1.3.1.3 Convolutional Neural Networks

When applied to images, most deep learning-based methods will utilize a Convolutional Neural Network (CNN) (Figure 1.14), which is a special type of ANN. In CNNs, a mathematical operation called a convolution, specifically designed for image processing, is used in place of the matrix multiplication typically performed in ANNs. Unlike fully connected layers, the convolutional layers in CNNs are not connected as densely, providing more flexibility and decreasing the number of weights per layer. This reduces some computational requirements without sacrificing the ability to derive features, shapes, and textures from images. Importantly, this feature extraction process occurs automatically through the training process, where the network determines identifying characteristics, representations, and features within the input.

![Figure 1.14: A CNN with a three-layered MLP to predict a single numerical output.](image)

CNNs are typically composed of multiple convolutional layers, which form the main building block of CNN architectures. Most of the computation within a CNN occurs in these layers through the convolution operation. Here, each of the layer’s convolution kernels iteratively moves across the image. At each iteration, the dot product is taken between the input and kernel. Over these iterations, the output from the series of dot products is built up into a feature map.

Each layer of a CNN is used to learn how to detect a different feature of the
input image. Because of this, the output feature map from each layer becomes the input for the subsequent layer. Through every subsequent layer, the kernels often become increasingly complex to better identify the unique features which represent the input image. Additional components such as pooling layers or MLPs may be present in the CNN to aid in reducing the number of parameters to be estimated, network complexity, and to perform classification on the extracted features to identify the input image, respectively.

1.3.2 Datasets and Data Considerations for Deep Learning

In deep learning, the dataset is typically divided into three parts: training, validation, and test sets. The training set is used to train the model by repeatedly adjusting the weights to minimize the loss function (see Section 1.3.1). The validation set is used to evaluate the performance of the model during training. Often, it is used to help tune the model’s hyperparameters, such as the learning rate. The validation set is typically used to select the best model among a set of models with different hyperparameters. The test set is used to evaluate the performance of the model after training is complete. The test set is used to assess performance metrics that are relevant to the task and can be used to ensure that the model does not overfit to the training set. Overfitting is a common problem in machine learning, where the model learns the training data too well and is unable to generalize to new, unseen data. Overfitting can occur when the model performs well in training, but poorly on new, unseen data. By holding out a portion of the dataset for validation and testing, the model can be evaluated on data that it has not seen during training, which helps to ensure that the model can generalize well and make accurate predictions on new data.

The use of training, validation, and test sets is important because the performance of deep learning models depends heavily on the available training data. Critically, in the absence of high quantities of high-quality data, models can perform sub-optimally. Larger datasets often produce better performance as they will have more examples from which the model can learn [97]. However, obtaining large datasets may be laborious, time-consuming, expensive, and simply not possible in
some cases, e.g. if a certain imaging technique is not widely adopted.

In the medical imaging domain, ethical and legal implications must be considered, including patient confidentiality, data security, and other applicable regulations [98]. Notably, other features of the data, such as any biases from the data selection, must also be carefully assessed. In some instances, the data sampling, populations, or annotation process may be biased. This can lead a model to produce inaccurate or ‘unfair’ predictions (i.e. those based on historical biases), which can have a variety of negative consequences [99–101]. These biases can yield diagnostic errors [99, 101] or amplify existing societal biases [99, 100], and as such, are important to be able to identify and mitigate.

Beyond considering the acquisition of large data, and the identification and mitigation of existing biases in those data, the data obtained must be of high quality. Poor-quality data, from either the input data or the ground-truth annotations and labels, can also negatively impact model performance by introducing errors into the training process.

Additional considerations must be made to ensure data diversity. Diverse data contain a range of examples taken from and which reflect real-world variability in attributes such as patient population, image modality, image quality, or disease. Using diverse data can help to control for changes in the distributions of training and testing data. In circumstances where the training and testing data have inherently different characteristics, poor generalization may result.

Data are a critical component of any deep learning method, which can greatly impact the performance, reliability, and effectiveness of a model. Careful consideration of the quality, size, diversity, and annotation of the dataset is important to ensure that the model can generalize well, make accurate predictions, and mitigate inherent biases. Poor-quality data, small datasets, biased datasets, and inadequate annotations can all lead to inaccurate or unfair predictions, while a diverse dataset with high-quality annotations can improve model robustness and generalization.
1.3.3 Deep Learning in Medical Imaging

One domain where deep learning and CNNs have had a large impact is medical image analysis [102–108]. Part of this impact stems from their ability to rapidly learn to optimize the kernels within their convolutional layers. This reduces the reliance on hand-engineered features and human intervention in image processing tasks such as disease prognosis and diagnosis, object detection, image segmentation, and image registration [102–108].

Computer-assisted disease prognosis and diagnosis can be an effective method of obtaining an additional opinion for a clinician, based on the information present in a patient’s images. Often, only a single diagnostic variable is to be determined; is the disease present or not present? While conventional computer-assisted systems leverage human-engineered features [109], deep learning methods have been able to assist in the differentiation of different types of lesions and determine disease progression and staging from various forms of imaging [102, 105, 106].

In many radiological workflows, the localization of lesions or anatomical structures is of critical importance. Traditionally, clinicians manually delineate different anatomical regions or structures. It follows that the ability to derive such annotations or detect objects automatically in a given image, as well as over time or between different patients, may be of great benefit to clinical practice to localize or track lesions, as well as for image discrimination (i.e. the identification of images that require further analysis or where a clinical decision has been indicated) [102, 106]. A step beyond identifying regions of interest with bounding boxes, or other arbitrarily-shaped geometric regions, is the generation of a pixel- or voxel-level segmentation for a 2D image or a 3D image volume. Image segmentation is the process of partitioning an image into multiple regions, each of which has a certain property or semantic meaning. In segmentation, each pixel or voxel is assigned a label for a pre-defined class, which can be used to identify the region of interest. This is done through the delineation of the boundaries of the region of interest, which can be used to generate the segmentation. These segmentations permit the quantitative analysis of images with respect to features such as shape and volume. When performed manually, this
can be a laborious and time-consuming process. Early automatic methods were based on discrete mathematical methods for digital signal/image processing, such as edge-detection filters. However, most recent approaches utilize deep learning methods. The applicability and effectiveness of deep learning-based segmentation methods have caused image segmentation to become one of the most common applications of deep learning in the medical imaging literature [106]. The most prevalent and widely used method for medical image segmentation, UNet [110], has been applied to countless different imaging modalities and anatomical regions [106]. While originally only intended for 2D images, extensions of UNet for 3D segmentation that are trained with 2D image slices [111] or 3D volumes [112] are now available. Not only do such methods often improve on the performance of conventional or hand-engineered methods, but they also typically require less time to compute the required segmentation [102, 106].

1.3.4 Deep Learning in Medical Image Registration

Deep-learning-based methods for medical image registration can be distinguished beyond the classifications of intensity- and feature-based, or rigid and non-rigid, given in Section 1.2.1. Given the inherent nature of network architectures, training processes (supervised, unsupervised, or otherwise), inference processes (one-shot or iterative), and output type(s) (parametric or non-parametric), among various defining characteristics, there are numerous other ways in which deep-learning-based registration methods may be categorized [103, 104, 107]. Notably, there are two widely-used approaches in the literature that have proven effective in obtaining highly-accurate registrations. The first method involves the use of deep learning to derive a similarity measure to be used in an iterative registration scheme (such as those described in Section 1.2.4), whereas the second method involves the direct prediction, often through regression, of the registration transformation itself, or, equivalently, the parameters which define this transformation. These approaches are described in the following subsections.
1.3.4.1 Deep Similarity Measures

The need to derive learned similarity measures comes from a natural extension of measures, such as MI, discussed in Section 1.2.3. In multimodal image registration, while hand-engineered measures were traditionally adopted, deep learning-based approaches have resulted in the development of novel learned similarity measures. The expectation for the effectiveness of such an approach has been based on the capabilities of deep-learning-based methods to outperform conventional solutions for many image classification and segmentation tasks.

Reframing the measurement of a similarity measure as a classification problem (e.g. as a measure of aligned or misaligned, as in [113]) or regression task (e.g. as a measure of similarity which may not be otherwise easily hand-engineered, as in [103, 104, 107]), it follows that learned similarity measures can take the place of conventional similarity measures when optimizing an iterative registration scheme. In several instances, these measures would be learned by training a network to classify the alignment of two input images as either aligned or not aligned, based on the image-pair-level label assigned to the image pair based on a predefined manual alignment. Other approaches sought the regression of objective, imaging-agnostic metrics, such as Target Registration Error (TRE), which is used widely to quantify registration accuracy relative to expert-aligned images as the reference standard [113]. With these learned measures, the trained network would subsequently produce the resulting similarity score or regressed metric during the iterative registration process to solve the registration problem [103, 104, 107].

1.3.4.2 Transformation Regression

The regression of a transformation or transformation parameters avoids the time-consuming iterative nature of a registration process to optimize a given similarity measure, regardless of whether this measure is learned or hand-engineered. However, solving the registration problem in this way brings new challenges: typically, ground-truth transformations are not available. This makes the use of supervised learning for deep learning-based registration methods non-trivial, or reliant on artificially-generated or simulated transformations. While it is possible to generate target
transformations from which the network may learn, such generated or simulated transformations may not accurately represent the desired ground-truth transformation, or consider biomechanical constraints appropriately within the context of a given application or domain. Importantly, this lack of ground-truth can lead to poor performance on real-world data if the generated or simulated transformation is too different from the true transformation [103, 104, 107].

The use of unsupervised learning and alternate forms of supervision, such as weakly-supervised learning, have permitted direct regression of the required transformation without a ground-truth transformation. With these methods, the transformed source image and the target image, or features derived from these images, may be used to directly regress a transformation with the learning guided by a computed similarity measure on the images or features [103,104,107]. Most unsupervised methods for medical image registration will compare the transformed source and target images directly with an intensity-based similarity measure, additionally applying some form of regularization in order to enforce smoothness on the predicted transformations or displacement field. In weakly-supervised methods, a higher level of correspondence is used, such as the segmentations of anatomical regions-of-interest on their own [114], or in tandem with the images [115], to compute a similarity measure based on the features and/or image intensities. Often, some form of regularization measure is applied as well.

While unsupervised and weakly-supervised registration methods do not often significantly outperform the state-of-the-art conventional iterative methods, they have been reported to perform required registrations much faster, while delivering comparable accuracy [103, 104, 107]. Beyond performance, rapid computation is also an important consideration for many clinical and interventional applications. In many instances, the reduction of computation time from, potentially, tens of minutes to sub-second can be essential for ensuring clinical translation may be viable given the reduction in the computational capacity required, as well as the elimination of time inconvenience to the end-user [103].
Medical Image Registration in Prostate Cancer Applications

Medical image registration is a core enabling technology for prostate cancer diagnosis and treatment through interventions such as MR-TRUS-guided, targeted prostate biopsy and brachytherapy, as discussed in Section 1.1.2. The use of registration to combine or spatially align or match high-quality, preoperative, diagnostic imaging, such as MR, to real-time, non-invasive imaging modalities, such as TRUS, has been investigated by many groups, and across many different application domains, including for prostate cancer interventions. Several existing works focus on the automatic rigid or non-rigid registration of MR to TRUS using intensity-, feature-, or learning-based methods [116, 117]. However, the intensity differences between MR and TRUS imaging makes accurate intensity-based registration difficult given the lack of detailed intraprostatic information and corresponding landmarks within the images themselves, especially within TRUS. This has led to much of the recent research in MR-TRUS registration in prostate cancer applications to focus on feature-based methods, and, more recently, learning-based methods [103, 118–120].

In conventional, iterative, intensity-based registration methods between MR and TRUS, MI is often used as a similarity measure. However, much of its use stems from variants that account for correlation-based ratios [121], or are conditioned contextually [122]. Normalized cross-correlation [123] and Modality Independent Neighbourhood Descriptors (MIND) [124] – as well as its multi-channel variant [125] – have also been investigated to account for the spatial and structural similarities between modalities, even when intensity characteristics differ greatly.

Looking to feature-based methods, the segmentation and subsequent extraction of relevant image features are, in general, always performed first. Subsequently, these features are registered. While not prostate or MR-TRUS specific, several common methods that have been widely applied to register extracted point-set or surface features, such as Iterative Closest Point (ICP) [126], Coherent Point Drift (CPD) [127], Thin-Plate Spline Robust Point Matching (TPSRPM) [128], or the application of a Gaussian mixture model (GMM) [129] for probabilistic point
correspondence prediction.

In practice, feature-based registration algorithms may require some form of deformable model in order to ensure that realistic and physically plausible deformations are accounted for [130–134]. Statistical deformation models which utilize surface points [135], or surface and intraprostatic points [130, 131] have been used for this purpose, whereas population-based models have also been developed to describe prostate motion in registration [132], which are then instantiated to generate a model for individual patients. Registration based on finite element model (FEM) deformation has also been used to constrain surface-based methods [133, 136].

Learning-based registration methods have demonstrated significant improvements over hand-engineered intensity- and feature-based methods, and are considered to be the state-of-the-art in terms of registration accuracy for many medical image registration tasks [103, 104, 107]. This is no different when considering the task of MR-TRUS registration. Owing to their ability to learn and infer deep similarity metrics and predict, or regress, the transformation or transformations directly through effective learning processes, as described in Sections 1.3.4.1 and 1.3.4.2, the prediction of similarity or transformation can often be performed in one or only a few forward passes through the model.

The application of learning-based methods to MR-TRUS registration has taken many forms, where inputs and training protocols may differ greatly. Several methods for learning deep similarity measures have been established [103, 104]. Notably, effective learning of similarity in terms of TRE in an adversarial manner has been demonstrated, where simulated deformation was added to previously manually-aligned images during the training process [137]. MR-TRUS-specific similarities can be learned directly from rigidly-aligned images. This permits improved registration accuracy when applied to conventional iterative registration processes – such as those described in Section 1.2.4 – as compared to “hand-engineered” measures, such as MI and MIND [113].

In practice, the direct regression of a transformation, either from images, features, or images and features together, is more commonly applied than the learning
1.4. Medical Image Registration in Prostate Cancer Applications

of a similarity measure in prostate MR-TRUS registration [103, 104]. Earlier deep-learning-based approaches utilized images more readily; Onofrey et al. [138] applied image synthesis to translate MR images into TRUS images. In doing so, this enabled the conversion of a multimodal registration problem into a unimodal one, where only the TRUS-TRUS registration must be performed, yet could still be applied to the MR images. The application of weak supervision, where image labels are used during training to compute a loss, but only the MR and TRUS images are needed at inference, has also demonstrated considerable success [114]. This approach has been shown to achieve high accuracy through a time and data-efficient inference process, where no separate segmentation is required [114]. Biomechanically-constrained point-set-based methods have also been learned, both from images [139] – where the deformation space is learned and applied as a constraint to the resulting deformation – and from FEM-based analysis of the prostate gland segmentations themselves [118].

In summary, though intensity-, feature- and learning-based methods may utilize different inputs, feature extraction processes, or compute different types of outputs, there are commonalities in approach and the challenges associated with each method. These challenges, with respect to their application in feature- and learning-based approaches (for both intensity- and feature-based learning approaches), are further examined in Section 3.1. Additionally, the application of different learning and training protocols and the use of user interaction (commonly required in many conventional image registration methods) are explored further in Sections 4.1 and 5.1.
Chapter 2

Motivation and Objectives

2.1 Clinical Motivations

The overarching clinical goal and motivation for medical image registration is to combine useful information present in two (or more) different images, from a single or different modalities, such that this information can be presented as a single image (or more generally, as a single visual representation). In doing so, the accuracy and reliability of diagnoses, surgical guidance, and a variety of other clinical decision-making processes or tasks may be improved or optimized. Having such registration algorithms and processes at the disposal of clinicians may reduce the cognitive load significantly, avoiding the necessity for a human operator to align image-derived information mentally, which is both a complex task and subject to significant inter- and intra-operator variability.

In the context of the application emphasized throughout this thesis, multimodal registration for enabling targeted prostate cancer biopsy guidance, there is an inherent desire to fuse the diagnostic and structural information available in MR imaging with real-time information available from TRUS imaging. Using both modalities in tandem may be critical to accurately and reliably performing effective biopsies. This is of importance in many real-world scenarios, such as when there is probe-induced tissue motion that may affect the registration. With real-time, adaptive methods, the impacts of unexpected tissue motion on the registration process may be lessened. Given the persistent issues with undersampling and underdiagnosis that can occur
during systematic biopsy procedures [18], the resulting action is often to acquire more samples [20] putting patients at higher risk for complications [18, 21]. This demonstrates a clear need to deliver simpler and more accurate methods for guidance that reduce cognitive load through rapidly employed, automatic, intraoperative registration and visualization.

The development of registration methods which permit more effective biopsies could permit the automation of a typically cognitive alignment process. During this process, where pre-operative MR is available, the optimal biopsy needle placements must be estimated based on separate (i.e. un-registered) MR and TRUS images. Therefore, with an effective and accurate registration, there is a potential for a reduction in cognitive load and subjectivity. Registration may permit faster, more accurate, and more consistent procedures with lower risk to patients, better sample localization and improved outcomes through lessened uncertainty and the targeting of a specific suspect or known cancerous regions in place of what is often referred to clinically as random prostate sampling. Similar benefits may be seen in prostate cancer therapies as well, such as those discussed in Section 1.1.2.3, where accurate registration can provide more accurate delivery of therapy, potentially reducing side effects due to localization errors or undertreatment, which is turn will improve patient outcomes.

While there is substantial evidence that supports the use of such registration-based image guidance methods to redefine and improve the standard of care for prostate biopsy [140–145], where improved outcomes are attained regardless of the clinician’s level of experience [143], the issues surrounding the cost-effectiveness [141], practicality [145], ease-of-use [142, 143, 145] and the possible requirement for additional training [142, 143, 145], remain. During an MR-TRUS fusion-guided targeted biopsy procedure, for example, each step of a typical TRUS volume acquisition, contouring (if required) and registration process can require several minutes [146]. In an already, relatively speaking, short procedure, typically taking approximately 30 minutes [147, 148], this additional time can have a substantial impact.

As a result, methods that deliver comparable or improved accuracy, but that
are also time- and data-efficient, may lessen the impact of the aforementioned concerns associated with adopting targeted protocols. Given the success that targeted biopsies have had, and their ability to successfully guide clinicians to improve biopsy outcomes and performance [149], efforts to reduce the difficulty in performing such procedures accurately may carry many additional downstream benefits for patient management. Thus, registration methods that require minimal time and can be performed during acquisition, instead of after, may be a critical step towards adopting image fusion-guided targeted biopsy approaches as the standard of care.

2.2 Thesis Objectives

The work described throughout this thesis has been performed with the aforementioned clinical motivations in mind, and with the underlying aim and objectives of developing novel deep-learning-based methods which permit fast, accurate, generalizable, yet data-efficient registrations. These methods are applied to scenarios in which the utilization of real-time imaging, such as US, is a key enabling technology. This includes the development of methods that:

1. Enable sparse, data-efficient, real-time interventional image registration;

2. Utilize human interaction to guide the learning of a patient-specific model;

3. Facilitate single-image-pair optimization for refining interpatient atlas generation, population-based assessments, as well as intrapatient longitudinal or change detection problems.

Given these objectives, the main contributions of this thesis are as follows:

1. Development and validation of a “model-free”, non-rigid method for general-purpose, unsupervised point-set registration of biomedical images;

2. Development and validation of a framework that intuitively combines error correction through user interaction with rapid adaptation through meta-learning and few-shot learning in order to optimize weakly-supervised biomedical image registration performance within a target domain;
Development and validation of a novel neural network training paradigm that combines conventional population-based, generalizable deep learning with single image pair optimization with a meta-learning approach.

2.3 Thesis Structure

The remaining chapters of this thesis are grouped into two main parts, where the major contribution and overarching methods presented in each part are summarized below.

Part 1 focuses on the non-rigid point-set registration problem as a solution for non-rigid multimodality image registration. In Chapter 3, a generalized framework for feature-based registration of biomedical images is presented, focusing on two applications where real-time registration processes are beneficial:

1. MR-TRUS registration, using complete and with partial/sparse TRUS image data for guiding prostate biopsy;

2. Spinal deformity quantification for scoliosis monitoring.

Part 2 explores the use of meta-learning-based approaches for prostate-imaging-related registration applications, first, for inference-time refinement and rapid adaptation of a real-time MR-TRUS registration process, with sparse imaging, which may be utilized to enable registration simultaneously during acquisition (Chapter 4). Then, in Chapter 5, meta-learning-based methods are applied to paired volumetric TRUS imaging, using an unsupervised approach for test-time optimization to improve inter-patient registration.

The final section of the thesis is a conclusion, which summarizes these aforementioned works in learning-based medical image registration and presents directions for future research.
Part I

Feature-based Multimodal Registration
Chapter 3

Learning Non-Rigid Partial Point-Set Registration

This chapter is based on the works entitled: “Multimodality biomedical image registration using free point transformer networks”, published in MICCAI ASMUS Workshop 2020 [119], “Real-time multimodal image registration with partial intra-operative point-set data”, published in Medical Image Analysis [120], and “Learning generalized non-rigid multimodal biomedical image registration from generic point-set data”, published in MICCAI ASMUS Workshop 2022 [150].

3.1 Introduction

Multimodal image registration is a subproblem of image registration wherein the images to be registered come from different scanner or sensor types. Multimodal registration methods have proven effective in image-guided interventions, where the aligned diagnostic information from preoperative imaging, such as MR or Computerized Tomography (CT), and intraoperative imaging, such as US, is displayed. This registered imaging aids in overcoming the typical restrictions and limitations of intra-operative imaging modalities, such as time-constraint, portability, ease of access, resolution, and field of view.

In many multimodal registration applications, such as MR-US or MR-CT alignment, intensity-based registration methods that minimize information-theoretic measures, such as mutual information and normalized mutual information [82, 124,
3.1. Introduction

or other statistical image similarity metrics [81, 82, 123, 155–158] have been widely investigated.

Despite the success of intensity-based registration methods, these can perform poorly for input image modalities with very different pixel/voxel intensity characteristics, such as MR and US. Most saliently, these differences often make it difficult to develop robust intensity-based registration methods that can generalize to different healthcare settings. In such cases, feature-based registration approaches provide a viable alternative for many clinical applications when features, such as organ boundaries, can be defined with minimal user interaction.

Feature-based methods have been widely employed within the field of medical imaging research not only for multimodal image registration methods but for registration in general [77, 159–161]. This is often due to their simpler and less computationally complex nature, with respect to intensity-based methods. Notably, many sparse, surface-point-set matching algorithms require some form of regularization. For example, statistical deformable models may be used to permit only physically plausible soft-tissue deformations [130–134, 162]. Additionally, the use of simple data formats, such as point-sets, can provide visually intuitive and easy-to-interpret representations of anatomical structures, which can aid clinical use and be an effective basis for clinical user interaction, such as manual refinement [163], providing feedback on registration uncertainty and quality [164].

Feature extraction has seen rapid advances in recent years given the development of automatic, well-validated, learning-based medical image segmentation methods. Such methods can yield real-time delineation of anatomical surfaces [110]. These surfaces may be sampled into point-sets for surface matching, for example, using classical point-set registration algorithms, such as ICP [126]. More contemporary alternatives, such as CPD [127] and TPSRPM [128] provide a solution for non-rigid registration. GMM have been used to compute registrations using probabilistic point correspondences [129].

Deep learning-based methods have gained traction in medical image registration, having made possible, or yielded improvement on, tasks thought previously infeasi-
ble without sufficient computational capacity. As such, making deep learning-based methods well suited to multimodal image registration problems. For consistency, the conventions surrounding intensity-based and feature-based registration defined in Section 1.2.1 are extended to deep-learning approaches. Therefore, feature-based deep learning registration methods are characterized as methods that utilize image features as network inputs (e.g. [118–120, 150, 165]), as opposed to using images as network inputs (e.g. [113, 114, 137]). Such feature-based methods may leverage existing, well-established, and well-validated automatic segmentation algorithms to, in some sense, convert the images into an image-acquisition-independent representation, required during inference. These representations shall have the same potential benefits, as those from the classical feature-based registration as argued above, to improve the generalization of developed methods by removing any artifacts or inconsistencies, and providing an intuitive and protocol- or scanner-independent form of the image.

CNNs have also been widely used to perform multimodal image registration. Some works approach the problem by learning similarity metrics directly from the images [113, 137], through image synthesis methods that convert the appearance of one or both input modalities such that they closely resemble the other before registration [138, 166, 167], or through reinforcement learning [168–170]. Image segmentation data may be used to learn non-rigid statistical deformation models which may, at inference, be used to guide a non-rigid surface registration [139]. Segmentations have been used to determine the correspondence between different imaging modalities in a weakly-supervised framework, with the advantage that the input images are only required at inference [114].

Despite the abundance of intensity-based deep learning methods described above, few methods utilize image features, such as surfaces, point-sets, or segmentations in addition to, or instead of, the images themselves. However, given the challenging nature of multimodality image registration, it may prove useful to leverage deep learning methods which can consume irregular data in addition to or instead of image data. The classical iterative methods for feature-based registration meth-
ods previously described [126–129] are not well-suited for applications requiring real-time registration since they are computationally intensive when processing large point/surface datasets. In contrast, the computationally efficient inference and the ability to model complex, nonlinear transformations of deep learning-based methods has motivated the development and application of neural networks to real-time registration [171–177]. Several such methods (for example, [171, 172]) have exploited PointNet [178], a deep learning framework for the classification and segmentation of point-sets as a feature-extraction method, precluding any learning of a registration process. One such method, PointNetLK [171], combined PointNet with the Lucas and Kanade algorithm to create an iterative, rigid point-set registration algorithm. Other works have applied PointNet as a means to learn hierarchical features to their method for 3D scene flow [172]. Other methods have been developed which do not use PointNet, such as those that provide iterative, self-supervised, rigid registration of partial point-sets with Partial Registration Networks (PRNet) [174], and rigid registration as a pre-cursor to ICP with Deep Closest Point [173]. Other approaches regressed correspondence between point-sets by using local and global features to compute the singular value decomposition for rigid registration [177].

It is notable that the above-mentioned methods only present rigid point-set registration, whereas for the purpose of image-guided interventions, non-rigid registration is more important. Non-rigid registration permits the fusion of imaging with spatial information and enables soft tissue motion compensation. Models which integrate local deformations have demonstrated improved registration in soft tissues by compensating for patient motion and other anatomical deformations [71]. Some non-rigid, point-set-based registration methods have also been proposed and applied to medical data. One for the analysis of lung motion, achieved by repurposing existing iterative algorithms within a deep learning framework [165]. A second provides multimodal image registration through the use of a deformation field defined by a weighted surface and volume point-set loss function through a matching process that explicitly models underlying biomechanical constraints [118]. Though these methods have shown promise, there have been limited subsequently published works
on deep learning-based multimodal image registration methods which are inherently feature-based. This may be explained by the difficulty associated with adequately selecting an effective parametric transformation model that does not explicitly define constraints on spatial coherence or smoothness. Additionally, these methods often rely on constrained models or need to explicitly model noise, outliers, and missing data. As such, these models may be inadequate to handle real-world data.

At inference, deep learning-based methods for multimodal image registration, be they intensity- or feature-based, typically require complete 3D volumetric data. In feature-based methods, this dependence on complete data makes the use of accurate and well-validated segmentation methods a necessity, as it is often considered impractical or infeasible for clinicians to manually segment all available images. When manual segmentations are performed, they are often inconsistent, error-prone and differ due to intra- and inter-observer variability [179]. In contrast, few deep learning-based registration methods explore the use of sparse or partially available data [171, 174] which is frequently encountered in surgical and interventional applications. However, these works produce only rigid or affine transformations, are not applied to medical image registration, and operate on sparsity or partial data availability in such a way that does not accurately reflect clinically realistic scenarios.

### 3.2 Contributions

This chapter describes a deep neural network architecture for non-rigid point-set registration, called FPT. The network consists of two parts: a global feature extraction module and a point transformation module. Importantly, FPT is not limited by the inherently unordered or only partially available structure of point-sets and predicts a non-rigid transformation that aligns them.

This work demonstrates the generalizability of FPT for point-set registration through a series of extensive experiments across multiple domains. First, robustness to noise, deformation, and missing data are demonstrated on the ModelNet40 dataset. Second, using FPT’s model-free approach and data-driven learning process, a pre-trained FPT is applied to the registration of real-world 3D reconstructions of
3.3 Methods

3.3.1 Free Point Transformer

Given a pair of source and target point-sets, \( \{ \mathbf{p}_s | s = 1, \ldots, N_s \} \) and \( \{ \mathbf{p}_t | t = 1, \ldots, N_t \} \), respectively, where \( \mathbf{p}_s \) and \( \mathbf{p}_t \) are \( D \)-dimensional vectors denoting individual point spatial coordinates in \( x \)-, \( y \)-, and \( z \) directions (here, \( D = 3 \)). The FPT framework aims to learn a model for inferring a transformation function \( T^{\{ \mathbf{p}_s \} \rightarrow \{ \mathbf{p}_t \}} \), between a pair of point-sets, such that it will map any new point, represented by the vector \( \mathbf{p}'_s \) in source coordinate space, to the target coordinates \( \mathbf{p}'_s \) as follows:

\[
\mathbf{p}'_s = T^{\{ \mathbf{p}_s \} \rightarrow \{ \mathbf{p}_t \}}(\mathbf{p}_s)
\]

(3.1)
where \( \tilde{p}_s \) is usually sampled from the source point-set domain, but not necessarily an element of \( \{ p_s \} \).

FPT models such a spatial point transformation using a parametric neural network \( T_{\theta}^{\{p_s\} \rightarrow \{p_t\}}(\tilde{p}_s) \), with network parameters \( \theta \), together with an end-to-end network training approach. The FPT network contains two modules: a global feature extraction module and a point transformation module. The global feature extraction module converts point-sets into a feature vector, whereas, the point transformation module predicts a displacement vector for the given input point \( \tilde{p}_s \) using the feature vector. A detailed illustration of the two modules and the network training scheme is shown in Figure 3.1. In the following sections, the construction and training details for these two modules, using a training set consisting of examples of different point-set pairs, is provided.

**Figure 3.1:** Schematic representation of the FPT network design for non-rigid point-set registration. The global feature extraction module takes a target and source point-set and applies shared input and feature transformations to both, creating a global feature vector. The point transformation module serves as a per-point transformation of the source point-set by determining the displacement to be added in order to obtain the transformed point-set.

Point-sets have important attributes which have been exploited in the design of the FPT, and which deliver several advantages for registration purposes: First, FPT accepts unordered and unstructured point-sets with a variable number of points. This requires the global feature extraction module to learn a representation, which determines a permutation, rotation, and cardinality invariant feature extraction step.
The global feature extraction module adapts the previously proposed PointNet architecture [178] to register a pair of point-sets. Second, the FPT has two separate functions: i) predicting a transformation from the network input, registration of the two input point-sets; ii) predicting displacements for individual given points.

These two functions are implemented with the global feature extraction module and the point transformation module, which are trained together but may be used for independent point-sets – i.e., those to register \( \{ \mathbf{p}_s \} \) and \( \{ \mathbf{p}_r \} \) – and those to be transformed \( \{ \mathbf{p}_t \} \). This flexibility is important as it allows the network inputs to be different from the point-sets used for computing the loss, which may only be available during training. Third, the point transformation module in the proposed FPT is defined without an explicit or parameterized registration method, permitting a “model-free” approach. This leads to non-rigid registration using a data-driven learning approach that prevents any collapse or folding that arguably may not be reflective of the data. FPT is trained without heuristic constraints, such as deformation smoothness or hand-engineered noise models. Training in this manner may ultimately be beneficial, as the restrictions imposed by such constraints or models enforce deformation that may over-simplify the complex soft tissue deformation and observable inter-structure motion that is possible with certain anatomical structures. As a result of these attributes and considerations, FPT is versatile and permits generalization to partial data while learning from complete data, as well as generalization to unseen types of objects. The FPT supports different types of learning supervision, including fully-supervised, semi-supervised, and partially- or weakly-supervised training see also a brief discussion in Section 3.3.2. However, in this work, the focus is placed on training the FPT using an unsupervised learning approach, which allows it to learn from raw point-set data without the need for ground-truth transformations. As is demonstrated through multiple applications and use-cases, permitting an end-to-end process to be achieved, which includes data acquisition followed by registration in real-time; an ability that is important in many tasks, such as time-critical medical applications of image registration.
3.3.1.1 Global Feature Extraction Module

PointNet [178] was originally designed to convert point-sets into permutation and rotation invariant feature vectors for classification and segmentation tasks. From the original PointNet architecture [178], this work has utilized the input and feature transformation and global information aggregation components to create high-dimensional feature vectors. Unlike the original PointNet architecture, which learns a $3 \times 3$ transformation matrix and subsequently multiplies this learned transform by the coordinates of the input points [178], FPT’s global feature extraction module learns a $4 \times 4$ transformation matrix to better allow for the representation of 3D translation in homogeneous coordinates, in addition to any rotation, scaling, shearing or reflections which may be represented in the original $3 \times 3$ transformation matrix. As in PointNet, this $4 \times 4$ transformation matrix is then used to transform the coordinates of the input points. This modification resulted from initial experimental results wherein a lower translational error was observed when the adapted PointNet was given the ability to encode translational differences more easily between point-sets in its feature representations. Additionally, batch normalization layers were removed from the PointNet to prevent the normalization of translational differences between source and target point-sets. In FPT, the above modifications create a single PointNet shared between the input point-sets $\{p_s\}$ and $\{p_t\}$, as illustrated in Figure 3.1. The module, in turn, generates feature vectors $g_s$ and $g_t$ with pre-defined lengths, from the source and target $\{p_s\}$ and $\{p_t\}$, respectively. These feature vectors are used to determine a global feature vector $g$:

$$g = f^{\{p_s\} \rightarrow \{p_t\}}_{\theta_{feat}} \tag{3.2}$$

where $g = [g_s^T, g_t^T]^T$ is the concatenated K-dimensional global feature vector and $f^{\{p_s\} \rightarrow \{p_t\}}_{\theta_{feat}}$ denotes the modified PointNet that represents a set-order-sensitive feature extraction function, i.e. $f^{\{p_s\} \rightarrow \{p_t\}}_{\theta_{feat}} \neq f^{\{p_t\} \rightarrow \{p_s\}}_{\theta_{feat}}$, which is invariant to the point-order in each set. $\theta_{feat}$ are the network parameters in the global feature extraction module.
3.3. Methods

3.3.1.2 Point Transformation Module

The point transformation module serves as a per-point transformation model \( f \) that predicts the displacement vector that transforms a point \( \tilde{p}_s \) to \( \tilde{p}_s' \), conditioned on the computed global feature vector \( g \) (given in Eq. 3.2):

\[
\tilde{p}_s' = f_{\theta_{trans}}(\tilde{p}_s \mid g)
\]  

(3.3)

In this work, a MLP network is used to model this transformation with network parameters \( \theta_{trans} \). Without loss of generality, the hidden units at \( l^{th} \) layer in an \( L \)-layer MLP, \( x^{(l)} = \left[ x_j^{(l)} \right]^T \), \( j = 1, \ldots, J^{(l)} \), representing the output feature vector with \( J^{(l)} \) \( (l = 1, \ldots, L) \) elements can be given in a recursive form:

\[
x_j^{(l)} = a^{(l)} \left( \sum_{j=1}^{J^{(l-1)}} w_j^{(l)} x_j^{(l-1)} + w_0^{(l)} \right)
\]  

(3.4)

where \( a^{(l)} \) is the element-wise activation function (rectified linear units are used in this work); and \( w_j^{(l)} (j = 1, \ldots, J^{(l-1)}) \) are the weights for each of the \( J^{(l-1)} \) elements in the input feature vector \( x^{(l-1)} = \left[ x_j^{(l-1)} \right]^T (j = 1, \ldots, J^{(l-1)}) \) from the previous layer. Together with the scalar bias weight \( w_0^{(l)} \), the point transformation module parameters are \( \theta_{trans} = \left[ \left[ w_j^{(l)} \right]^T \right]_{j=0, 1, \ldots, J^{(l-1)}}^{l=1, \ldots, L} \).

The point transformation module \( f_{\theta_{trans}} \) is specified by the module input and output, the point-concatenated global feature vector \( x^{(0)} = [g^T, \tilde{p}_s^T]^T \) and the displacement vector \( d_i = x^{(L)} \), respectively; therefore, \( J^{(0)} = K + 3 \) and \( J^{(L)} = 3 \). The transformed point can be computed by \( \tilde{p}_s' = \tilde{p}_s + d_i \). Predicting the displacement \( d_i \), instead of the transformed points \( \tilde{p}_s' \) directly, which was found empirically to simplify the initialization for model training. It is important to note that the transformation model parameterized by the above-described MLP does not have constraints on the transformation smoothness, which are commonly imposed with assumptions such as coherence between adjacent points, giving a less constrained transformation.

The use of MLP parameterization also facilitates an efficient one-dimensional (1D) convolution implementation for multiple feature vectors during the FPT network training. For each network input point-set pair, \( \{p_s\} \) and \( \{p_t\} \), the global feature
extraction module computes a global feature vector $g$ (Eq. 3.2). In the general case, the point transformation module aims to transform a point-set $\{\mathbf{p}_s\}, \ s = 1, \ldots, N_s$, using Eq. 3.3, conditioned on the same global feature vector $g$. Assume a row-wise concatenated “feature matrix” $\mathbf{M}^{(l)}, \ l = 1, \ldots, L$, at $l^{th}$ layer, such that:

$$
\mathbf{M}^{(l)} = \begin{bmatrix}
(x^{(l)}_{s=1})^T \\
(x^{(l)}_{s=2})^T \\
\vdots \\
(x^{(l)}_{s=N_s})^T
\end{bmatrix}, \text{ where } \mathbf{M}^{(0)} = \begin{bmatrix}
g^T, \mathbf{p}_{s=1}^T \\
g^T, \mathbf{p}_{s=2}^T \\
\vdots \\
g^T, \mathbf{p}_{s=N_s}^T
\end{bmatrix} \text{ and } \mathbf{M}^{(L)} = \begin{bmatrix}
d^T_{i=1} \\
d^T_{i=2} \\
\vdots \\
d^T_{i=N_s}
\end{bmatrix}
$$

Now, for computing the output feature vector at the $l^{th}$ layer, substituting the network weight $w_j^{(l)}$ in Eq. 3.4 with a scalar weight $k_j^{(l)}$, the $j^{th}$ of the $J^{(l-1)} \times 1$D convolution kernels for each of the $J^{(l)}$ elements. The $J^{(l-1)} \times J^{(l)}$ kernels are convolved over all $N_s$ elements in the column space of the feature matrices $\mathbf{M}^{(l)}$ because the MLP weights are shared between all the input row vectors $[g^T, \mathbf{p}_s^T]$ in the feature matrices. The rows representing different feature-vector-concatenated points in $\{\mathbf{p}_s\}$ remain independently multiplied by the 1D kernel.

### 3.3.2 FPT Network Training

The FPT network described here was trained to optimize the network parameters $\theta = [\theta^{T}_{\text{feat}}, \theta^{T}_{\text{trans}}]^T$ by minimising the distance between the transformed source point-set $\{\mathbf{p}_s^{\prime}\}$ and a given target point-set $\{\mathbf{p}_t\}, \ t = 1, \ldots, N_t$. The specific form of the function $\mathcal{L}(\{\mathbf{p}_s^{\prime}\}, \{\mathbf{p}_t\} | \theta)$ serves as the training loss and is described in Section 3.3.2.1, while the goal of the network training is:

$$
\hat{\theta} = \arg\min_{\theta} \mathbb{E}_{\Omega} \left[ \mathbb{E}_{\Omega} \left[ \mathcal{L}(\{\mathbf{p}_s^{\prime}\}, \{\mathbf{p}_t\}) \right] \right]
$$

(3.6)

where substituting the transformed source point-set $\{\mathbf{p}_s^{\prime}\}$ as:

$$
\hat{\theta} = \arg\min_{\theta} \mathbb{E}_{\Omega} \left[ \mathbb{E}_{\Omega} \left[ \mathcal{L}(\{\mathcal{T}_{\theta}(\mathbf{p}_s^{\prime})\rightarrow\{\mathbf{p}_t\}), \{\mathbf{p}_t\}) \right] \right]
$$

(3.7)
3.3. Methods

where \( \widetilde{p}_t' = T_{\theta_{\text{trans}}} (\widetilde{p}_s) = f_{\theta_{\text{trans}}} (\widetilde{p}_s) f_{\theta_{\text{feat}}} (\widetilde{p}_s) \) is parameterized by two neural networks, as described above; \( \mathbb{E}_\Omega [\cdot] \) and \( \mathbb{E}_e [\cdot] \) denote the expectation operators over the training point-set-pair domains \( \Omega \) and \( e \), training examples for \( \{p_s\} \) and \( \{p_t\} \) to compute the global feature, and training examples of \( \{ \hat{p}_s \} \) and \( \{ \hat{p}_t \} \) for computing the distance-based loss, respectively. This general form of training lets FPT provide the flexibility to allow the network input point-sets \( \{p_s\} \) and \( \{p_t\} \) to differ from the training “ground-truth” point-sets \( \{ \hat{p}_s \} \) and \( \{ \hat{p}_t \} \). This formulation can be applied in the following scenarios:

1. Unsupervised learning of point-set registration, i.e., \( \{ \hat{p}_s \} = \{p_s\} \) and \( \{ \hat{p}_t \} = \{p_t\} \).

2. Partial data registration with full data available in training, e.g., \( \{p_s\} \subseteq \{ \hat{p}_s \} \) or \( \{p_t\} \subseteq \{ \hat{p}_t \} \). This will provide a loss computed from, in general, stronger supervision \( \{ \hat{p}_s \} \) and \( \{ \hat{p}_t \} \), while test data at inference are more likely to have a different distribution that is similar to what is represented by \( \{p_s\} \) and \( \{p_t\} \).

3. Training-time bootstrap resampling [180], when \( \{ \hat{p}_s \}, \{ p_s \}, \{ \hat{p}_t \} \) or \( \{ p_t \} \) is large or the difference between their sizes – i.e., the difference in the number of points – is large. This allows sampling a subset of any of these point-sets during a stochastic or mini-batch gradient descent while maintaining an unbiased gradient.

4. Weakly-supervised learning, i.e., \( \{p_s\} \supseteq \{ \hat{p}_s \} \) or \( \{p_t\} \supseteq \{ \hat{p}_t \} \), yet \( \{ \hat{p}_s \} \) and \( \{ \hat{p}_t \} \) are point-sets with known point-to-point correspondence, which are available during training, while \( \{p_s\} \) and \( \{p_t\} \) are the network input available during inference.

In the remainder of this chapter, several of these formulated scenarios are explored. In Section 3.4, Scenario 1 and Scenario 2 are demonstrated on a general-purpose computer vision dataset, with simulated transformations – unknown to the network and training process – and occluded data. In Section 3.5, Scenario 2 is demonstrated using pre-trained models applied to a challenging spinal curvature
quantification problem. Notably, training FPT with the data used in this section would present a use-case for the formulation described in Scenario 3 given the large nature of the source and target point-sets. In Section 3.6, Scenario 1 and Scenario 2 are demonstrated on a prostate MR to TRUS registration problem with full and partial data. The use of known landmarks (such as lesions, calcifications, or other anatomical structures) for known point correspondence in training during this problem, though not presented in this work, would represent the formulation presented in Scenario 4.

3.3.2.1 Loss Functions

In this work, two loss functions are compared to measure the distance between two point-sets. The first is a widely used metric for determining the mean nearest neighbor distances between point-sets; the $D_C$ [181]. Second, a negative log-likelihood of a GMM is employed to encourage the network to minimize the difference between the distributions of the point-sets. A two-way $D_C$ is used in this work as follows:

$$\mathcal{L}_{D_C}(\{\tilde{p}_s\}, \{\tilde{p}_t\}) = \frac{1}{N_t} \left( \sum_{r \in [1, N_t]} \min_{s \in [1, N_s]} ||\tilde{p}_r - \tilde{p}_s'||^2 \right) + \frac{1}{N_s} \left( \sum_{s \in [1, N_s]} \min_{r \in [1, N_t]} ||\tilde{p}_s - \tilde{p}_r'||^2 \right)$$

(3.8)

Unlike $D_C$, the negative log-likelihood of a GMM is one of the alternatives which requires explicit parameters when considering outliers or noise levels. It is assumed that the spatially transformed source point-set $\{\tilde{p}_s\}$ defines the centers of the $\tilde{N}_s$ Gaussian clusters in a mixture model with $\tilde{N}_s + 1$ clusters, with the additional cluster being a uniform distribution (with a probability of $\frac{1}{\tilde{N}_t}$) for potential outliers [127]. Given the target point-set $\{\tilde{p}_t\}$ as the model-fitting data, the GMM training loss can then be defined as the negative log-likelihood function of the mixture model, as follows:
\[ \mathcal{L}_{GMM}(\{\mathbf{p}_s^t\}, \{\mathbf{p}_t\}) = -\sum_{t=1}^{N_t} \log \left( \frac{1}{N_t} \left[ \frac{1}{N_s} \sum_{s=1}^{N_s} \left( e^{-\frac{\|\mathbf{p}_t - \mathbf{p}_s^t\|^2}{2\sigma^2}} \right) \right] \right) \] (3.9)

where \( \sigma^2 \) is the isotropic covariance; \( \frac{1}{N_s} \) is the equal membership probabilities among the \( N_s \) Gaussian clusters that are defined by the source point-set; and \( u, 0 \leq u \leq 1 \), weights the uniform distribution. The loss function thus has two parameters, \( \sigma^2 \) and \( u \). As the FPT architecture does not explicitly constrain transformations from potential folding, collapse, or severe distortion of the transformed point-sets, the two-way construction of the loss functions, including both the \( D_C \) and negative log-likelihood of a GMM, are employed in this work. It is interesting to find that, during the training, relaxing constraints such as ‘one-to-one’ correspondence did not cause unrealistic, extremely non-smooth deformation.

### 3.4 Simulated Registrations on Computer Vision Datasets

The rigid and non-rigid registration performance of FPT is presented with added noise, deformation, and missing data. In these experiments, the Princeton ModelNet40 dataset [182] is used. The use of ModelNet40 permits a simple comparison between FPT and existing classical and learning-based methods to determine if the framework is sound and effective for general-purpose rigid and non-rigid point-set registration ahead of utilizing it for more challenging biomedical applications, where smaller datasets and more varied data distributions are common.

#### 3.4.1 Data

The Princeton ModelNet40 dataset, which consists of 12,311 geometric surface models of objects spanning 40 categories, split object-wise into a 9,843 object training set and 2,468 object testing set [182]. ModelNet40 is a commonly used collection of 3D models of the most common categories in the world for computer vision classification, segmentation, and registration methods. The input point-set for
FPT comprised 2,048 points sampled from the surfaces of ModelNet40 shapes. The points were used for the target and source point-sets in training and testing.

### 3.4.2 Network Implementation and Training

The previously described FPT was implemented in TensorFlow [183] and Keras [184]. The FPT network architecture used in the below experiments uses a value of $K = 1024$; thus, extracting a 1024-dimensional feature vector from each of the twin weight-sharing PointNets. In the point transformer module, $L = 6$, and $U = \{1024, 512, 256, 128, 64, 3\}$. The set of activation functions, $a$, was defined such that the first five layers used the ReLU activation function [96], but the final layer used no activation function (i.e. a linear activation).

All networks were trained for 2,000 epochs with the Adam optimizer [185], a minibatch size of 32, and an initial learning rate of $1 \cdot 10^{-3}$ when training. Networks were trained on an NVIDIA DGX-1 system using a single Tesla V100 Graphics Processing Unit (GPU).

During training, the points in $P_T$ and $P_S$ were permuted, scaled, and deformed on-the-fly. The points were scaled, per-sample, between $[-1, 1]$ in each of the $X$, $Y$ and $Z$ directions. $P_S$ was further transformed on-the-fly with deformation, rotation, and displacement. Rotation angles were randomly sampled from $[-45^\circ, 45^\circ]$ about each of the $X$, $Y$ and $Z$ axes. Displacements were randomly sampled from $[-1, 1]$ in each of the $X$, $Y$ and $Z$ directions. Non-rigid deformation was simulated by a radial basis function (RBF) deformation model with a Gaussian kernel. RBF deformation was defined by a perturbation of the control points by Gaussian random shift ($\mu = 0, \sigma = 0.1$). The scaled, deformed, and transformed version of the input was used as the source point-set.

Given the number of training samples and iterations required in training, computational efficiency was important when selecting a deformation model. A RBF was used as it produces smooth deformations and is computationally efficient enough to be applied on-the-fly during training and prior to inference during testing. Experiments with other deformation models, such as elastic body spline models [186], are needed to assess FPT’s ability to reconstruct more localized displacements.
When training FPT for partial registration, occlusion was simulated by selecting a random point on the model surface and discarding the 512 (25% of the input points) nearest neighbor points. Following point removal, translation, rotation, and deformation were performed as described above.

3.4.3 Comparison with Classical and Learning-Based Methods

It is of note that FPT predicts point displacements with one pass through the network and no subsequent refinement. However, the classical and learning-based methods that are compared to in the below-described experiments solve registration in various ways. ICP [126] and PointNetLK [171] compute the rigid transformation matrix directly. CorsNet [177] and PRNet [174] predict point correspondence and compute the optimal rigid transformation through a least-squares fit. Some methods are applied iteratively [127, 174] whilst others preclude refinement [171]. No existing methods provide partial and deformable registration, focusing instead on only one of the problems in isolation.

Unseen Objects. FPT was evaluated on rigid and non-rigid transformations. Transformations are generated as in training, although for the rigid transformations, no non-rigid deformation was added. As such, the inherent design and training of FPT is unchanged - meaning it may predict a transformation that contains deformation to the input point-sets, but the ground-truth transform between the input point-sets does not have any deformation added for the rigid transformation registrations. In this experiment, FPT is compared to rigid iterative and learning-based methods, as well as non-rigid iterative methods.

Unseen Objects with Gaussian Noise. FPT was evaluated with Gaussian noise introduced to the source point-set. Non-rigid transformations are generated as in training, with Gaussian noise of variable standard deviation (SD) added to simulate noisy data.

Unseen Objects with Varying Deformation. FPT was evaluated while increasing deformation in the source point-set. Non-rigid transformations are generated as in training, though variable deformation was introduced by modifying the SD of the distribution from which perturbations of the control points which defined the RBF
were drawn.

**Partial Registration of Unseen Objects.** FPT’s performance was evaluated for partial registration. Transformations are generated as in training for the ‘partial-to-full’ registrations, where point removal is only performed on the source point-set. For ‘partial-to-partial’ registrations, point removal is performed on both the source and target point-sets. FPT is compared to rigid iterative and learning-based methods. Notably, one such method [174] is designed explicitly for partial registration, whereas FPT is not. In training FPT for this experiment, only the input data is modified, as in Section 3.4.2, and not the architecture. This change allows assessment of the effectiveness of FPT’s architecture on a broad range of registration problems wherein the input is not two complete point-sets.

### 3.4.4 Evaluation Methods

The accuracy of the registrations was evaluated by measuring the root mean square error (RMSE) of the rotational (R) and translational (t) errors. Angular measurements are given in degrees. The accuracy of the point-level registration errors is given by the $D_C$ and the Hausdorff Distance ($D_H$). As the $D_C$ was used as the loss function for all experiments, it serves as a useful indicator of network generalization on test data, while the $D_H$ gives a worst estimate of the distances between the point-sets. Additionally, the computational time was also recorded for each registration experiment.

### 3.4.5 Results & Discussion

**Unseen Objects.** Figure 3.2 shows FPT’s registration performance with rigid and non-rigid input transformations. Table 3.1 gives registration performance of FPT and other methods on unseen objects. FPT outperformed all other evaluated methods with respect to rotational error. FPT was comparable or superior to other methods with respect to translational error when presented with rigid and non-rigid input transformations. FPT matched or outperformed other learning-based methods with respect to inference time, and was over 50 times faster than CPD - an established and widely used iterative non-rigid registration method. Additionally, FPT performed
3.4. Simulated Registrations on Computer Vision Datasets

consistently in both rigid and non-rigid registration, demonstrating its ability to effectively perform both types of registration.

Figure 3.2: Rigid (top) and non-rigid (bottom) registrations with FPT. Blue: source, yellow: target, green: transformed source.

Table 3.1: Registration performance of FPT and other methods on unseen objects in the ModelNet40 dataset. The lowest mean value in each section is bolded.

<table>
<thead>
<tr>
<th>Method</th>
<th>Transformation</th>
<th>Time (s)</th>
<th>RMSE (R)</th>
<th>RMSE (t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICP [126]</td>
<td>Rigid</td>
<td>0.05</td>
<td>28.84(^\circ)</td>
<td>0.193</td>
</tr>
<tr>
<td>PointNetLK [171]</td>
<td>Rigid</td>
<td>0.14</td>
<td>14.47(^\circ)</td>
<td>0.045</td>
</tr>
<tr>
<td>CorsNet [177]</td>
<td>Rigid</td>
<td>0.08</td>
<td>16.24(^\circ)</td>
<td>0.012</td>
</tr>
<tr>
<td>CPD [127]</td>
<td>Rigid</td>
<td>5.94</td>
<td>8.29(^\circ)</td>
<td>0.049</td>
</tr>
<tr>
<td>CPD [127]</td>
<td>Non-Rigid</td>
<td>6.01</td>
<td>8.39(^\circ)</td>
<td>0.051</td>
</tr>
<tr>
<td>FPT</td>
<td>Rigid</td>
<td>0.08</td>
<td>5.01(^\circ)</td>
<td>0.015</td>
</tr>
<tr>
<td>FPT</td>
<td>Non-Rigid</td>
<td>0.08</td>
<td>5.18(^\circ)</td>
<td>0.032</td>
</tr>
</tbody>
</table>

Unseen Objects with Gaussian Noise. Figure 3.3a shows FPT’s non-rigid registration performance with Gaussian noise. Table 3.2 gives non-rigid registration performance of FPT with Gaussian noise. The rotational and translational errors were not largely affected by the introduction of Gaussian noise to the source point-set. Further, D_C and D_H metrics, as computed on the registration performed with the training set, remained fairly constant until larger amounts of Gaussian noise are introduced. This suggests that outlier points have the most substantial effect on the registration of those outlier points; however, the overall registration quality, as
measured by RMSE, was not largely unaffected by noise, likely owing to the fact that FPT extracts global features.

![Figure 3.3: Non-rigid registration with added Gaussian noise (a) and varying deformation (b) using FPT. Gaussian noise added at 0.01 SD (left), 0.02 SD (center), 0.04 SD (right). Deformation added at 0.1 SD (left), 0.2 SD (center), 0.4 SD (right). Blue: source, yellow: target, green: transformed source.](image)

### Table 3.2: Non-rigid registration performance on unseen objects with added Gaussian noise.

<table>
<thead>
<tr>
<th>Noise (SD)</th>
<th>RMSE (R)</th>
<th>RMSE (t)</th>
<th>D_C</th>
<th>D_H</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5.17°</td>
<td>0.032</td>
<td>0.029</td>
<td>0.101</td>
</tr>
<tr>
<td>0.005</td>
<td>5.47°</td>
<td>0.032</td>
<td>0.030</td>
<td>0.103</td>
</tr>
<tr>
<td>0.01</td>
<td>5.40°</td>
<td>0.033</td>
<td>0.031</td>
<td>0.108</td>
</tr>
<tr>
<td>0.02</td>
<td>5.55°</td>
<td>0.034</td>
<td>0.035</td>
<td>0.123</td>
</tr>
<tr>
<td>0.04</td>
<td>5.59°</td>
<td>0.037</td>
<td>0.042</td>
<td>0.167</td>
</tr>
</tbody>
</table>

### Unseen Objects with Varying Deformation. Figure 3.3b shows FPT’s registration performance at varying amounts of deformation in the non-rigid input transformations. Table 3.3 gives the registration performance of FPT in the deformation experiments. FPT’s rotational and translational errors, D_C, and D_H, as computed by an FPT registration performed with the training set, was more noticeably affected when deformation was introduced with σ = 0.2. This suggests that larger deformations have a greater effect on the overall registration, as FPT was not able to as successfully reconstruct the original shape - as in Figure 3.3b. As FPT was trained with perturbation added at σ = 0.1, it is suspected that introducing larger deformations in training may improve FPT’s ability to perform successful registrations with greater deformation.

### Partial Registration of Unseen Objects. Figure 3.4 shows FPT’s ‘partial-to-full’ and ‘partial-to-partial’ non-rigid registration performance. Table 3.4 gives registration
performance for FPT and other evaluated methods for ‘partial-to-full’ and ‘partial-to-partial’ registrations. Here, it is seen that FPT was comparable or superior to other learning-based methods. While PRNet maintains superior performance in both rotational and translation error, it is explicitly designed for partial registration and iteratively refines the predicted registration [174]. Furthermore, the use of a modified $D_C$ limits FPT’s ability to perform partial registration as it relies on the existence of a one-to-one correspondence between point-sets. When points are removed from the source or target point-set, as in partial registration, a two-way $D_C$ will compute distances in a one-to-many manner, as some points will be distant from the other point-set. However, this limitation may be alleviated through the formulation of a one-way $D_C$ loss for partial registration applications. As such, FPT’s performance in partial registration must be further validated with other losses and training protocols in future work.

Table 3.4: Partial registration performance of FPT and other methods on unseen objects.

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Transformation</th>
<th>RMSE (R)</th>
<th>RMSE (t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICP [126]</td>
<td>Partial-to-partial</td>
<td>Rigid</td>
<td>32.40°</td>
<td>0.279</td>
</tr>
<tr>
<td>PointNetLK [171]</td>
<td>Partial-to-partial</td>
<td>Rigid</td>
<td>16.58°</td>
<td>0.048</td>
</tr>
<tr>
<td>PRNet [174]</td>
<td>Partial-to-partial</td>
<td>Rigid</td>
<td>3.20°</td>
<td>0.016</td>
</tr>
<tr>
<td>FPT</td>
<td>Partial-to-partial</td>
<td>Rigid</td>
<td>6.97°</td>
<td>0.063</td>
</tr>
<tr>
<td>FPT</td>
<td>Partial-to-partial</td>
<td>Non-Rigid</td>
<td>7.99°</td>
<td>0.082</td>
</tr>
<tr>
<td>FPT</td>
<td>Partial-to-full</td>
<td>Rigid</td>
<td>5.79°</td>
<td>0.053</td>
</tr>
<tr>
<td>FPT</td>
<td>Partial-to-full</td>
<td>Non-Rigid</td>
<td>6.34°</td>
<td>0.068</td>
</tr>
</tbody>
</table>
3.5 Quantification of Scoliotic Spine Curvature

As a preliminary real-world use-case for generalized non-rigid point-set registration, FPT’s effectiveness in the quantification of spinal curvature for scoliosis measurement is investigated. Scoliosis is a spinal deformity identified in 3% of adolescent children [187]. Scoliosis is often monitored and measured with X-ray imaging (X-ray), however, repeated use of X-ray has been linked to an increased incidence of cancer [188]. US imaging has been proposed as a safer, more accessible option for scoliosis monitoring and measurement [189–191]. Recent deep learning methods for automatic bone segmentation in US have been shown to adequately reconstruct the overall spinal curvature in pediatric patients with scoliosis [191]. However, compared to X-ray, these reconstructions are not sufficient without additional visual context. This is due in part to the lack of some meaningful features, depending on the orientation of the images used to acquire the images used in the reconstruction, which can support clinical decision-making and create clear visualizations for patients. The registration of an overall shape or curvature to a generic spine model stands to provide appropriate and meaningful spine visualization. Manual registration of such reconstructions to generic spine models for visualization and measurement is possible, though the process is time-consuming, error-prone, and operator dependant [190]. These limitations reveal the need for fast, automatic
3.5. Quantification of Scoliotic Spine Curvature

methods which register US-based reconstructions to generic anatomical models for use in place of or to supplement traditional medical imaging.

3.5.1 Data

Using Ungi et al.’s method for automatic bone segmentation from US [191], 13 different curvatures from 7 different patients were reconstructed. The surfaces of the reconstructions, as well as a generic spine model (Figure 3.5) were resampled into point-sets, and subsequently registered to evaluate FPT’s ability to generalize to unseen shapes.

3.5.2 Network Implementation

To demonstrate FPT’s generalizability to unseen objects from outside the domain of the training set, the implementation of FPT used for this experiment was not fine-tuned with any data from this application. Instead, the resulting trained network from the experiments in Partial Registration of Unseen Objects, given in Section 3.4.3, was used directly.

3.5.3 Evaluation Method

Spinal curvature is commonly quantified by the Cobb angle - the angle formed between lines drawn on the end-plates of vertebrae above and below the main curvature [192]. However, as vertebral end-plates are not visible in US, they do not appear in the 3D reconstructions. As such, spinal curvature is reported using the transverse process angle (TxA) as it is visible in X-ray and US, and has a very strong correlation to Cobb angle [189]. TxA is defined by the angle formed between lines drawn through the lateral ends of each transverse process above and below the main curvature [189]. TxAs were calculated on the X-ray and the deformed spine model. The reported TxA in the X-ray was defined by two lines in 2D. The TxA in the deformed model was defined by two lines in 3D, with the reported TxA being that which was computed by projecting these lines into 2D in the coronal plane. The lines annotated on the X-ray and the deformed model were drawn by a doctor of chiropractic medicine with over 10 years of experience in scoliosis measurement in X-ray and over 5 years of experience with US.
3.5.4 Results & Discussion

TxA were measured in both the X-ray and in the deformed spine models to permit comparison between the clinical standard for scoliosis measurement and FPT's registration. The 13 spinal curvatures acquired from X-ray in the dataset measured between $6.4^\circ$ and $11.5^\circ$. The maximum difference between TxA measurements from X-ray and the deformed model was $2.3^\circ$. The average difference between TxA measurements from X-ray and the deformed model was $1.3 \pm 0.8^\circ$. Figure 3.6 graphically
3.5. Quantification of Scoliotic Spine Curvature

demonstrates the aforementioned results. Figure 3.7 illustrates an anterior-posterior and lateral visualization of a representative case from the dataset.

![Scatter plot of per-patient TxA measured from the deformed spine models vs. X-ray-based measurements. Figure from [150].](image)

**Figure 3.6:** Scatter plot of per-patient TxA measured from the deformed spine models vs. X-ray-based measurements. Figure from [150].

In practice, only curvatures larger than 5° are considered clinically significant. Additionally, the curvature from X-ray measurements may vary by up to 5° due to inter-observer variability and the time of day at which the images are acquired [193]. As such, for monitoring and measuring scoliosis, an accuracy within 5–10° is considered acceptable and serves its purpose for determining the next steps and course of care for a given patient. While this proof-of-concept experiment is limited by the patient sample size, and the scale of the curvatures, given the availability of paired US imaging and corresponding X-ray, it is clear that the results presented are clinically acceptable. Without any fine-tuning, and having provided FPT with geometries that are external to the training set, FPT was able to register the spine models with an average error of less than 1.5°. Importantly, all measured differences fell within the 5° clinical error range, permitting a promising future use for the
3.6 Prostate MR-TRUS Registration

In this section, prostate MR-TRUS registration is employed as an additional example that demonstrates how FPT may be applied for real-world clinical applications. Prostate MR-TRUS image fusion is a technique for using MR images to perform tumor-targeted needle biopsy [194, 195] and minimally-invasive treatments [196] in patients for whom clinically significant prostate cancer is suspected or confirmed. The techniques involve presenting information on the location and size of MR-visible lesions/tumors to complement the information provided by real-time TRUS images so that needles and other instruments can be placed to accurately target specific tissue.

Figure 3.7: Sample anterior-posterior (top) and lateral (bottom) visualization from patient data. Left to right: X-ray image, 3D reconstruction, deformed spine model, deformed spine model overlaid on 3D reconstruction, deformed spine model overlaid on the X-ray image. Solid lines show measurements in X-ray images, dashed lines show measurements on deformed spine models. Figure from [150].

creation of accurate 3D visualizations that may be used for monitoring and measuring scoliosis without the need for error-prone manual registration processes, or the use of X-ray imaging and its associated ionizing radiation.
regions. MR-derived lesion/tumor information is typically displayed as a visual overlay superimposed on TRUS images, as a composite MR-TRUS image, or with the MR and TRUS images presented side-by-side. Displaying the images using any of these methods requires accurate registration.

3.6.1 Data

The experimental dataset used in the evaluation comprised 108 pairs of pre-operative T2-weighted MR and intraoperative TRUS images from 76 patients which were acquired during the SmartTarget clinical trials [146]. A mixture of 1.5T and 3T MR scanners at University College London Hospital were used for the T2-weighted MR acquisition. A standard clinical TRUS machine equipped with a bi-plane transducer probe at University College London Hospital was used for TRUS acquisition. Parasagittal slices were acquired and reconstructed into a 3D volume based on the interpolated TRUS intensity values at measured locations across a rectangular grid. Before point-set generation from the prostate contours, each of the MR and TRUS images was resampled to an isotropic voxel size of $0.8 \times 0.8 \times 0.8$ mm$^3$. Prostate gland boundaries were segmented in the resampled MR and TRUS images. Segmentations of the prostate gland in the MR images were acquired as part of the SmartTarget clinical trial protocols [146]. Additionally, segmentations of the prostate gland in the TRUS images were manually edited by biomedical imaging researchers based on automatically contoured prostate glands from the original TRUS slices [197]. Each annotator had at least 5 years of experience in MR-TRUS registration, for manual segmentation of both image types. One additional annotator with over 10 years of experience performed a final quality control check for all cases. All annotators had previously completed a two-day advanced course for radiologists, hosted by the Urology Department and the Radiology Department at University College London Hospitals, on the use and interpretation of MR and TRUS images for prostate cancer diagnosis and treatment.

Using segmented MR and TRUS images, the contours and volumes of each prostate gland were extracted to generate two 3D point-sets – $P_T$ from the TRUS images and $P_S$ from the MR images – using a grid-based sampling approach in which
3.6. Prostate MR-TRUS Registration

Each voxel was converted into a vector of its \([x, y, z]^T\) 3D Euclidean coordinates in the segmented image volume. When the points are displayed, this gives the appearance of a grid-like point-set. As such, each voxel’s location represented a single point in space in the generated point-set (Figure 3.8).

![Figure 3.8](image)

**Figure 3.8:** Point-sets depicting (from left to right) the anterior, right, posterior, and left views of a prostate volume obtained from a segmented TRUS (top) and MR volume (bottom) for one pair of patient images. Figure from [120].

3.6.2 Network Implementation and Training

This series of experiments was performed using the afore-described implementation of FPT given in Section 3.4.2. As previously, the FPT network architecture was employed with a value of \(K = 1024\). In the point transformer module, \(L = 6\), and \(U = \{1024, 512, 256, 128, 64, 3\}\). The set of activation functions, \(a\), was defined such that the first five layers used the ReLU activation function [96], but the final layer used no activation function (i.e., a linear activation). An illustration of FPT network, as applied for prostate MR-TRUS fusion is presented in Figure 3.9.

All networks were trained for 2,000 epochs with the Adam optimizer [185], a minibatch size of 8, and an initial learning rate of \(1 \cdot 10^{-5}\) when training with the \(D_C\) Loss (“FPT-Chamfer”). When training with the GMM Loss (“FPT-GMM”), all hyperparameters were identical to those used when training with the \(D_C\) Loss, apart from a minibatch size of 2 due to larger memory requirements for the computation of the loss. Additionally, hyperparameters \(\sigma^2\) and \(u\) in the GMM Loss were set as
Figure 3.9: Schematic representation of the FPT network design for non-rigid point-set registration as applied to the MR-TRUS fusion task. Figure from [120].

0.001 and 0.1, respectively. Networks were trained on an NVIDIA DGX-1 system using a single Tesla V100 GPU.

During training, the points in $P_T$ and $P_S$ were permuted, scaled, and resampled on-the-fly. The points were scaled, per-sample, between $[-1, 1]$ in each of the $X$, $Y$ and $Z$ directions. Both point-sets were then shuffled and randomly subsampled to the desired cardinality. $P_S$ was further transformed on-the-fly with scaling, rotation, and displacement. Rotation angles were randomly sampled from $[-45^\circ, 45^\circ]$ about each of the $X$, $Y$ and $Z$ axes. Displacements were randomly sampled from $[-1, 1]$ in each of the $X$, $Y$ and $Z$ directions.

3.6.3 Gaussian Radial Basis Function Network

To demonstrate the effectiveness of FPT, it is compared to the use of a parametric transformation, similar to that proposed by [118]. This network uses a Gaussian radial basis function (G-RBF) model to compute, and account for, the complex deformation between the surfaces of the source and target point-sets. In the implementation of the G-RBF network, the global feature extraction module developed for FPT is used, however, the point transformation module is replaced with a G-RBF module. This G-RBF module determines the point displacements by predicting the control
points ($X_{\text{Control}}$) and spline coefficients ($X_{\text{Coefficients}}$) required to compute the G-RBF transformation. Subsequently, the input point-sets are registered by non-rigid transformation via computation of the displacement between source and target point clouds, providing the transformed source points as:

$$\tilde{\mathbf{p}}_s' = f_{\theta_{G-RBF}}(\tilde{\mathbf{p}}_s) = \mathbf{\alpha} \mathbf{k}_s + \tilde{\mathbf{p}}_s$$

(3.10)

where $\mathbf{k}_s = [k_1^s, k_2^s, \ldots, k_{N_c}^s]^T$ is the $N_c \times 1$ Gaussian kernel vector $k_c^s(\mathbf{p}_c, \tilde{\mathbf{p}}_s) = e^{-\frac{||\mathbf{p}_c - \tilde{\mathbf{p}}_s||^2}{2\sigma^2}}$ [198], computed for each point $\tilde{\mathbf{p}}_s$, with respect to a set of control points $\mathbf{p}_c$, $c = 1, \ldots, N_c$. The $N_c$ control points $\mathbf{p}_c$ and the $3 \times N_c$ spline coefficients $\mathbf{\alpha}$ are directly predicted by the G-RBF point transformation network. In this work, the G-RBF uses $N_c = 27$ control points and a kernel parameter $\sigma = 1$, unless otherwise indicated. The G-RBF network was trained with the same amount of training data, the same data augmentation methods, and the same loss functions as the FPT.

Two variants of the proposed G-RBF network were compared, each with different loss functions used in training: First, using the $D_C$ Loss (“G-RBF-Chamfer”), and secondly, using the (“G-RBF-GMM”), both as previously described in Section 3.3.2.1. An illustration of the G-RBF network, including the G-RBF point transformation module is presented in Figure 3.10

### 3.6.4 Comparison with G-RBF and Classical Methods

In addition to the previously described G-RBF networks, the FPT was compared to two example pairwise iterative methods for point-set registration: the widely used rigid ICP algorithm and the non-rigid CPD algorithm. In the experiments, the ICP algorithm was permitted to complete up to 25 iterations. All other parameters and initializations were performed as described by Besl and McKay [126]. For the CPD algorithm, $w = 0$, where the value of $w$ ($0 \leq w \leq 1$) indicates the assumed amount of noise present in the point-set and permitted the algorithms to complete up to 250 iterations. All other parameters were set to the default values described by Myronenko and Song [127].

A series of experiments were performed to demonstrate the FPT’s performance
Figure 3.10: Schematic representation of the G-RBF network design for non-rigid point-set registration. Similarly to FPT, the global feature extraction module takes a target and source point-set and applies shared input and feature transformations to both, creating a global feature vector. The G-RBF point transformation module applies a non-rigid transformation using the predicted control points and spline coefficients, as in Eq. 3.10, to obtain the transformed point-set. Figure from [120].

compared to the G-RBF networks (G-RBF-Chamfer and G-RBF-GMM), ICP, and CPD for registration of MR to TRUS data. In these experiments, the same dataset described above (see Section 3.6.1) is utilized, comprising 108 pairs of MR and TRUS images. These data were split into a training and testing set, wherein 70% of the data (75 MR-TRUS pairs) were reserved for training, and 30% of the data (33 MR-TRUS pairs) were reserved for testing. Any patient who had multiple series of imaging was assigned to the training set to ensure that images from a single patient were not included in both the training and the test set. A hold-out set was not used to prevent bias by an exhaustive hyperparameter search when creating and training FPT to demonstrate the ability of its data-driven architecture compared to other methods. It should be noted that this two-way split experiment may systematically underestimate the registration performance of the G-RBF network and other methods which rely on extensive hyper-parameter tuning.
MR to TRUS Registration. The non-rigid registration performance of FPT when aligning complete volumetric MR and TRUS point-sets was evaluated similarly to the first scenario presented in Section 3.3.2. Performance in this registration problem was tested by varying the size of \( \{p_s\} \) and \( \{\tilde{p}_s\} \). Both the FPT and the G-RBF network were trained using both loss functions, using 1,024, 2,048, 4,096, or 8,192 points in \( \{p_s\} \). Owing to the cardinality invariance of the FPT and G-RBF network architectures, each of these trained networks was then used to predict registrations with 1,024, 2,048, 4,096, or 8,192 points in \( \{p_s\} \) to test the sensitivity to different sampling rates between network inputs during training and at inference. The ICP and CPD algorithms do not require training and were evaluated on the computed registrations they produced with 1,024, 2,048, 4,096, or 8,192 input points.

MR to Partial TRUS Registration. Additionally, the non-rigid registration performance of FPT when aligning complete volumetric MR point-sets to partial volumetric TRUS point-sets was evaluated similarly to the second scenario presented in Section 3.3.2. This series of experiments was designed to reflect three clinical scenarios in which only point-data defining part of the prostate surface are available due to a small number of 2D TRUS images being acquired, each representing a different slice through the organ. For each of these scenarios, surface points extracted from only two or three segmented ultrasound slices were used as inputs to the registration algorithms, reducing the amount of available data considerably.

Examples of each prostate TRUS imaging scenario investigated are illustrated in Figure 3.11. The first scenario represents the case where three evenly distributed 2D TRUS slices are obtained. The second scenario represents the case where two or three TRUS slices are obtained, but the slices are biased to one lateral direction. The third scenario represents the case where only two TRUS slices are obtained which provide poor coverage of the prostate gland, with the slices skewed to the left or right side.

To quantitatively describe and validate the differences between each scenario, two metrics were defined and used: ‘slice centroid distance’ and ‘slice span’. The slice centroid distance was defined as follows:
3.6. Prostate MR-TRUS Registration

Figure 3.11: Illustration of possible TRUS images acquired from the TRUS transducer. Images are captured in the sagittal plane (left) and are shown with other image slices that may be acquired in each of the three scenarios, from an axial view (right). Figure from [120].

\[
\text{Slice Centroid Distance} = \|c_p - c_s\|_2
\]  
(3.11)

where \(c_p\) is the geometric center of the TRUS prostate point-set, and \(c_s\) is the geometric center of the point-set comprising all the selected TRUS image slices. Additionally, the slice span was defined as follows:

\[
\text{Slice Span} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|c_p - c_{is}^i\|_2^2}
\]  
(3.12)

where the set \(\{c_{is}^i \mid i = 1, \ldots, n\}\) describes the centroid points comprising the individual selected TRUS image slices from \(n\) slices. These metrics are illustrated in Figure 3.12 and expected and computed values for the metrics which quantitatively describe the distribution of points and individual frames in each scenario are given in Table 3.5.

In the first set of experiments, changing the point sampling rates in training and at inference is seen to not affect the selected registration metrics (see Section 3.6.4). Therefore, in this second set of experiments, only instances of FPT-Chamfer and G-RBF-Chamfer with input point-sets containing 4,096 points in each of the three TRUS
Table 3.5: Qualitative metrics which describe the three clinical scenarios. Values: Mean ± SD mm

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Slice Centroid Distance</th>
<th>Slice Span</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expected</td>
<td>Actual</td>
</tr>
<tr>
<td>Lowest</td>
<td>4.07 ± 0.94</td>
<td>Highest</td>
</tr>
<tr>
<td>Between S1 &amp; S3</td>
<td>4.78 ± 1.43</td>
<td>Between S1 &amp; S3</td>
</tr>
<tr>
<td>Highest</td>
<td>10.6 ± 2.63</td>
<td>Lowest</td>
</tr>
</tbody>
</table>

scenarios were trained. An input size of 4,096 points was selected empirically as the previous experiment demonstrated no clear difference in registration quality when varying input point-set sizes. The $D_C$ Loss was selected as it reduced training time and produced superior results for registration error when compared to FPT networks trained with the GMM Loss in the previous experiment. For additional comparison, the ICP and CPD algorithms were also evaluated in these three previously described scenarios with input point-sets containing 4,096 points.
3.6.5 Evaluation Methods

The accuracy of the prostate surface point registrations was quantified using the \(D_C\), the \(D_H\), and TRE, calculated as the distance between points defining the 3D locations of corresponding, manually identified anatomical landmarks in the TRUS and MR images \([114, 199]\). The \(D_C\) was used as the loss function for some of the experiments and therefore indicates the network generalization to independent test data. Together with the \(D_H\), the registration accuracy on the point-set-represented individual prostate glands can be measured. The TRE is defined as the root-mean-square of each of the distances computed between the geometric centroids of the registered pairs of source and target landmarks for each patient. The landmarks consisted of 309 pairs of points, including points defining the apex, base, urethra, visible lesions, junctions between the gland, gland zonal separations, vas deferens and the seminal vesicles, and other patient-specific point landmarks such as calcifications and fluid-filled cysts \([146]\). It is noted that while the term ‘landmark’ is often used to denote a ‘point landmark’, here it is used to define a pixel-wise label of the structure in the image. As such, for non-point or non-spherical structures, the geometric centroid of the structure is used as the landmark for the purpose of computing TRE. Additionally, it should be noted that the overall spatial distribution of these landmarks may be representative of the full TRE distribution in this application \([114, 118–120, 123–125, 130–133, 137, 139, 154, 157, 162, 167, 169, 170, 200–207]\), but landmark-based TREs nevertheless provide a useful estimate of the errors associated with localizing tumors within the prostate. The computational time was also recorded for each registration experiment.

3.6.6 Results & Discussion

3.6.6.1 MR-TRUS Registration

Table 3.6 shows the mean and SD for \(D_C\), \(D_H\), and TRE for each of the different methods and input point-set sizes. A Shapiro-Wilk test was performed on the values for \(D_C\), \(D_H\), and TRE of each method at each point-set size. In all instances, the test did not show evidence of non-normality \((p > 0.05)\). Across all variants and
3.6. Prostate MR-TRUS Registration

experiments in MR to TRUS registration, FPT-Chamfer achieves a mean TRE of 4.71 mm, compared to 5.16 mm for FPT-GMM, 5.29 mm for G-RBF-Chamfer, 5.25 mm for G-RBF-GMM, 6.02 mm for ICP, and 5.08 mm for CPD. Without any form of alignment on the dataset, TRE of 25.43 mm is observed. FPT-Chamfer gives the lowest average $D_C$ and TRE in nearly all instances, while CPD gives the lowest average for $D_H$. The prostate contours from a sample slice in the transverse plane from resulting registrations of three cases with each of the learning-based and iterative methods are shown in Figure 3.13.

Table 3.6: $D_C$, $D_H$, and TRE for each method used and at each point-set size in the first MR-TRUS registration experiment. The lowest mean value in each section is bolded. Significant differences with respect to FPT-Chamfer are denoted with an asterisk (*), based on two-tailed paired t-tests at $\alpha = 0.05$. Values: Mean ± SD mm.

<table>
<thead>
<tr>
<th>Points in ${\mathbf{p}_i}$</th>
<th>Method</th>
<th>$D_C$</th>
<th>$D_H$</th>
<th>TRE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPT-Chamfer</td>
<td>1.10 ± 0.17</td>
<td>6.18 ± 1.36</td>
<td>4.75 ± 1.45</td>
</tr>
<tr>
<td></td>
<td>FPT-GMM</td>
<td>1.14 ± 0.18</td>
<td>6.65 ± 1.53*</td>
<td>5.49 ± 1.68*</td>
</tr>
<tr>
<td></td>
<td>G-RBF-Chamfer</td>
<td>1.15 ± 0.17</td>
<td>7.20 ± 1.43*</td>
<td>5.50 ± 1.61*</td>
</tr>
<tr>
<td></td>
<td>G-RBF-GMM</td>
<td>1.17 ± 0.16*</td>
<td>7.70 ± 2.00*</td>
<td>4.87 ± 1.33</td>
</tr>
<tr>
<td></td>
<td>ICP [126]</td>
<td>1.29 ± 0.19*</td>
<td>8.48 ± 1.79*</td>
<td>5.94 ± 1.68*</td>
</tr>
<tr>
<td></td>
<td>CPD [127]</td>
<td>1.25 ± 0.19*</td>
<td>6.12 ± 1.32</td>
<td>5.12 ± 1.35</td>
</tr>
<tr>
<td></td>
<td>FPT-Chamfer</td>
<td>1.38 ± 0.19</td>
<td>6.27 ± 1.41</td>
<td>4.69 ± 1.41</td>
</tr>
<tr>
<td></td>
<td>FPT-GMM</td>
<td>1.41 ± 0.20</td>
<td>6.66 ± 1.36*</td>
<td>5.34 ± 1.50*</td>
</tr>
<tr>
<td></td>
<td>G-RBF-Chamfer</td>
<td>1.42 ± 0.19</td>
<td>7.30 ± 1.48*</td>
<td>4.92 ± 1.34</td>
</tr>
<tr>
<td></td>
<td>G-RBF-GMM</td>
<td>1.45 ± 0.21*</td>
<td>7.83 ± 1.66*</td>
<td>5.19 ± 1.46*</td>
</tr>
<tr>
<td></td>
<td>ICP [126]</td>
<td>1.53 ± 0.22*</td>
<td>8.58 ± 1.86*</td>
<td>6.03 ± 1.62*</td>
</tr>
<tr>
<td></td>
<td>CPD [127]</td>
<td>1.44 ± 0.20</td>
<td>5.98 ± 1.34</td>
<td>4.99 ± 1.39</td>
</tr>
<tr>
<td></td>
<td>FPT-Chamfer</td>
<td>1.72 ± 0.21</td>
<td>6.46 ± 1.20</td>
<td>4.55 ± 1.34</td>
</tr>
<tr>
<td></td>
<td>FPT-GMM</td>
<td>1.75 ± 0.23</td>
<td>6.81 ± 1.32</td>
<td>4.70 ± 1.52</td>
</tr>
<tr>
<td></td>
<td>G-RBF-Chamfer</td>
<td>1.77 ± 0.23*</td>
<td>7.50 ± 1.56*</td>
<td>5.42 ± 1.75*</td>
</tr>
<tr>
<td></td>
<td>G-RBF-GMM</td>
<td>1.80 ± 0.24*</td>
<td>7.94 ± 1.78*</td>
<td>5.52 ± 1.40*</td>
</tr>
<tr>
<td></td>
<td>ICP [126]</td>
<td>1.89 ± 0.23*</td>
<td>8.63 ± 1.84*</td>
<td>6.09 ± 1.53*</td>
</tr>
<tr>
<td></td>
<td>CPD [127]</td>
<td>1.73 ± 0.22</td>
<td>6.07 ± 1.13*</td>
<td>4.98 ± 1.42*</td>
</tr>
<tr>
<td></td>
<td>FPT-Chamfer</td>
<td>2.16 ± 0.29</td>
<td>6.73 ± 1.22</td>
<td>4.83 ± 1.42</td>
</tr>
<tr>
<td></td>
<td>FPT-GMM</td>
<td>2.19 ± 0.29*</td>
<td>7.10 ± 1.30</td>
<td>5.09 ± 1.40</td>
</tr>
<tr>
<td></td>
<td>G-RBF-Chamfer</td>
<td>2.20 ± 0.30*</td>
<td>7.82 ± 1.56*</td>
<td>5.34 ± 1.64*</td>
</tr>
<tr>
<td></td>
<td>G-RBF-GMM</td>
<td>2.24 ± 0.30*</td>
<td>8.16 ± 1.79*</td>
<td>5.40 ± 1.89*</td>
</tr>
<tr>
<td></td>
<td>ICP [126]</td>
<td>2.34 ± 0.33*</td>
<td>9.17 ± 2.03*</td>
<td>6.01 ± 1.79*</td>
</tr>
<tr>
<td></td>
<td>CPD [127]</td>
<td>2.10 ± 0.25*</td>
<td>6.49 ± 1.42</td>
<td>5.21 ± 1.34*</td>
</tr>
</tbody>
</table>

For the FPT-Chamfer and FPT-GMM implementations, between 14 and 50
3.6. Prostate MR-TRUS Registration

Figure 3.13: Example image slices illustrating registration results from three different cases, shown in the transverse plane. Each image shows the original TRUS image with the transformed source (MR) contours (red) superimposed onto the target (TRUS) contours (green). Columns indicate registrations from each method; from left to right: FPT-Chamfer, FPT-GMM, G-RBF-Chamfer, G-RBF-GMM, ICP, and CPD. Figure from [120].

registrations may be performed per second, depending on the size of the input point-set at inference. These registration times are nearly identical to those achieved with the G-RBF network (G-RBF-Chamfer and G-RBF-GMM), approximately 5–8 times faster than those observed with ICP, and approximately 200–5,000 times faster than those observed with CPD. Table 3.7 shows the mean and SD of the registration times for the different methods.

Table 3.7: Time to compute a single registration at a given point-set size for FPT, G-RBF, ICP and CPD. Values: Mean ± SD s

<table>
<thead>
<tr>
<th>Points in {\hat{\mathbf{p}}_s}</th>
<th>1,024</th>
<th>2,048</th>
<th>4,096</th>
<th>8,192</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPT</td>
<td>0.02 ± 0.00</td>
<td>0.02 ± 0.01</td>
<td>0.04 ± 0.01</td>
<td>0.07 ± 0.01</td>
</tr>
<tr>
<td>G-RBF</td>
<td>0.02 ± 0.00</td>
<td>0.02 ± 0.01</td>
<td>0.04 ± 0.01</td>
<td>0.07 ± 0.01</td>
</tr>
<tr>
<td>ICP [126]</td>
<td>0.10 ± 0.01</td>
<td>0.15 ± 0.02</td>
<td>0.31 ± 0.02</td>
<td>0.45 ± 0.02</td>
</tr>
<tr>
<td>CPD [127]</td>
<td>4.17 ± 1.04</td>
<td>17.89 ± 3.15</td>
<td>85.58 ± 10.52</td>
<td>357.73 ± 25.74</td>
</tr>
</tbody>
</table>

To assess if the changes in \(D_C\) and \(D_H\) when using different numbers of points at inference is related to the size and inherent point density of \(\{\hat{\mathbf{p}}_s\}\), the results of MR to TRUS registrations performed on a grouped series of random subsamples without replacement is also reported. By creating multiple unique and non-intersecting
subsets of points for each MR and TRUS prostate volume, each of the predicted registrations may be grouped, and subsequently combined with, their accompanying subsets. This was performed at four different thresholds for the size of \( \{ \tilde{p}_r \} \); in essence, eight registrations were performed on eight subsets of 1,024 points, four registrations were performed on four subsets of 2,048 points, two registrations were performed on two subsets of 4,096 points, and one registration was performed on the original 8,192 points. The results of these grouped registrations are presented with the mean and SD of \( D_C \), \( D_H \), and TRE at the four different thresholds in Table 3.8.

Table 3.8: \( D_C \), \( D_H \), and TRE for each method used and at each grouped registration threshold. Values: Mean ± SD mm

<table>
<thead>
<tr>
<th>Points in ( { \tilde{p}_r } )</th>
<th>Groups</th>
<th>Method</th>
<th>( D_C )</th>
<th>( D_H )</th>
<th>TRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>8,192</td>
<td>1</td>
<td>FPT-Chamfer</td>
<td>1.10 ± 0.17</td>
<td>6.03 ± 1.35</td>
<td>4.80 ± 1.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FPT-GMM</td>
<td>1.16 ± 0.18</td>
<td>6.82 ± 1.49</td>
<td>5.33 ± 1.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G-RBF-Chamfer</td>
<td>1.15 ± 0.18</td>
<td>7.33 ± 1.43</td>
<td>5.43 ± 1.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G-RBF-GMM</td>
<td>1.18 ± 0.17</td>
<td>7.88 ± 1.99</td>
<td>5.58 ± 1.40</td>
</tr>
<tr>
<td>4,096</td>
<td>2</td>
<td>FPT-Chamfer</td>
<td>1.10 ± 0.17</td>
<td>6.20 ± 1.34</td>
<td>4.74 ± 1.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FPT-GMM</td>
<td>1.12 ± 0.16</td>
<td>6.60 ± 1.47</td>
<td>5.37 ± 1.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G-RBF-Chamfer</td>
<td>1.14 ± 0.16</td>
<td>7.21 ± 1.45</td>
<td>5.21 ± 1.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G-RBF-GMM</td>
<td>1.17 ± 0.19</td>
<td>7.76 ± 2.04</td>
<td>5.12 ± 1.58</td>
</tr>
<tr>
<td>2,048</td>
<td>4</td>
<td>FPT-Chamfer</td>
<td>1.10 ± 0.16</td>
<td>6.37 ± 1.36</td>
<td>4.67 ± 1.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FPT-GMM</td>
<td>1.12 ± 0.17</td>
<td>6.30 ± 1.48</td>
<td>4.94 ± 1.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G-RBF-Chamfer</td>
<td>1.13 ± 0.19</td>
<td>7.07 ± 1.46</td>
<td>5.28 ± 1.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G-RBF-GMM</td>
<td>1.16 ± 0.18</td>
<td>7.57 ± 2.01</td>
<td>5.35 ± 1.68</td>
</tr>
<tr>
<td>1,024</td>
<td>8</td>
<td>FPT-Chamfer</td>
<td>1.09 ± 0.17</td>
<td>6.00 ± 1.33</td>
<td>4.79 ± 1.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FPT-GMM</td>
<td>1.11 ± 0.18</td>
<td>6.31 ± 1.50</td>
<td>5.23 ± 1.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G-RBF-Chamfer</td>
<td>1.15 ± 0.19</td>
<td>7.23 ± 1.39</td>
<td>5.18 ± 1.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G-RBF-GMM</td>
<td>1.16 ± 0.18</td>
<td>7.55 ± 1.96</td>
<td>5.21 ± 1.56</td>
</tr>
</tbody>
</table>

It is observed that grouping registrations with different point sampling rates at inference does not appear to affect \( D_C \) or \( D_H \). This demonstrates that differences in \( D_C \) and \( D_H \) in prior experiments may be due to point-set density; where a less dense point-set produces a higher value for the same metrics. It may be concluded that, with sufficient points sampled in each set, the obtained TRE became less sensitive to the tested different sampling strategies and increase of the sampling density, a practically desirable property of the proposed method. While there are small variations in the average reported TRE for the grouped registrations, FPT-Chamfer still produces the
lowest overall average TRE in the registrations performed at each threshold for the size of \( \{\hat{p}_e\} \).

### 3.6.6.2 MR-Partial TRUS Registration

Table 3.9 shows the mean and standard deviation for \( D_C \), \( D_H \) and TRE for each of the different methods in each different clinical scenario. A Shapiro-Wilk test was performed on the values for \( D_C \), \( D_H \), and TRE of each method in each scenario. In all instances, the test did not show evidence of non-normality \((p > 0.05)\). Across all methods and scenarios in the MR to partial TRUS registration, FPT-Chamfer achieves the lowest average \( D_C \), \( D_H \) and TRE in all instances. Among the deep learning-based methods, average \( D_C \), \( D_H \), and TRE are similar to those in the first series of experiments where complete data were available and \( \{\hat{p}_e\} \) contained 4,096 points. ICP and CPD demonstrate lower average performance in all metrics relative to their results in the previous experiment. Most saliently, \( D_C \) and \( D_H \) for CPD are 1.5–6 times higher on average than in the previous experiment, and computed values of TRE nearly double on average. Based on two-tailed paired t-tests at \( \alpha = 0.05 \), the differences in \( D_H \) and TRE across all three scenarios between FPT-Chamfer and G-RBF-Chamfer, FPT-Chamfer and ICP, and FPT-Chamfer and CPD are statistically significant \((p < 0.005, p < 0.005, p < 0.001, \text{respectively})\). The differences in \( D_C \) across all three scenarios between FPT-Chamfer and ICP, and FPT-Chamfer and CPD are also statistically significant \((p < 0.005, \text{and } p < 0.001, \text{respectively})\). 3D visualizations of the prostate shapes before and after registration with variants of FPT for three different cases are given in Figure 3.14. The prostate contours from a sample slice in the transverse plane from resulting registrations of three cases with each of the scenarios for FPT-Chamfer are shown in Figure 3.15. A box plot of the TREs at comparing FPT-Chamfer, G-RBF-Chamfer, ICP, and CPD at the patient level for MR to TRUS and MR to partial TRUS registrations in all three scenarios is given in Figure 3.16.

### 3.6.6.3 Discussion

Unlike intensity-based methods, wherein similarity metrics are often utilized, the FPT leverages the geometric and spatial information from point-sets to drive the
### Table 3.9: $D_C$, $D_H$, and TRE for each method used in the partial MR-TRUS registration experiment. All point-sets are of size 4,096. The lowest mean value in each section is bolded. Significant differences with respect to FPT-Chamfer are denoted with an asterisk (*), based on two-tailed paired t-tests at $\alpha = 0.05$. Values: Mean ± SD mm.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Method</th>
<th>$D_C$</th>
<th>$D_H$</th>
<th>TRE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPT-Chamfer</td>
<td>1.40 ± 0.20</td>
<td>6.38 ± 1.48</td>
<td>4.88 ± 1.56</td>
</tr>
<tr>
<td></td>
<td>G-RBF-Chamfer</td>
<td>1.45 ± 0.20</td>
<td>7.38 ± 1.68*</td>
<td>5.39 ± 1.79*</td>
</tr>
<tr>
<td></td>
<td>ICP [126]</td>
<td>1.93 ± 0.41*</td>
<td>9.07 ± 1.95*</td>
<td>6.54 ± 2.19*</td>
</tr>
<tr>
<td></td>
<td>CPD [127]</td>
<td>2.25 ± 0.42*</td>
<td>15.74 ± 4.29*</td>
<td>9.35 ± 3.04*</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>FPT-Chamfer</td>
<td>1.41 ± 0.21</td>
<td>6.36 ± 1.70</td>
<td>4.81 ± 1.75</td>
</tr>
<tr>
<td></td>
<td>G-RBF-Chamfer</td>
<td>1.46 ± 0.22</td>
<td>7.68 ± 1.73*</td>
<td>5.27 ± 1.95*</td>
</tr>
<tr>
<td></td>
<td>ICP [126]</td>
<td>1.94 ± 0.43*</td>
<td>9.28 ± 2.24*</td>
<td>6.48 ± 2.24*</td>
</tr>
<tr>
<td></td>
<td>CPD [127]</td>
<td>3.32 ± 0.72*</td>
<td>21.38 ± 6.28*</td>
<td>9.69 ± 3.36*</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>FPT-Chamfer</td>
<td>1.42 ± 0.21</td>
<td>6.62 ± 1.90</td>
<td>4.76 ± 1.71</td>
</tr>
<tr>
<td></td>
<td>G-RBF-Chamfer</td>
<td>1.45 ± 0.23</td>
<td>7.71 ± 2.11*</td>
<td>5.55 ± 2.38*</td>
</tr>
<tr>
<td></td>
<td>ICP [126]</td>
<td>1.94 ± 0.62*</td>
<td>9.12 ± 2.59*</td>
<td>7.04 ± 2.33*</td>
</tr>
<tr>
<td></td>
<td>CPD [127]</td>
<td>6.58 ± 1.03*</td>
<td>36.78 ± 7.27*</td>
<td>10.30 ± 3.74*</td>
</tr>
</tbody>
</table>

Learning and subsequent registration process. While the effectiveness of this work relies on the extraction of features from the imaging data, the point-sets required may be generated efficiently and automatically via accurate image segmentations obtained from emerging deep learning methods [118, 197]. For MR-TRUS registration, only a few 2D US images may need to be segmented to produce a sufficient number of input points for registration using the aforementioned grid-based sampling approach described in this work. Furthermore, using FPT-Chamfer, TREs are lower or comparable to all other methods, with a mean TRE in the first and second experiments of 4.71 mm and 4.81 mm, respectively. As illustrated in Table 3.9, FPT-Chamfer significantly outperforms other methods in the partial registration, in all metrics, except for G-RBF-Chamfer, with a two-way $D_C$ loss (Eq. 3.8), when evaluating also using $D_C$. Other independent metrics, including $D_H$ and TRE, have all supported the superior generalization ability from FPT, with statistical significance.

Previous work comprehensively reports registration accuracies for full data sets, i.e., full 3D prostate gland segmentations may reduce variance in registration error given the results observed in this work, although additional validation is needed to
3.6. Prostate MR-TRUS Registration

Figure 3.14: MR-TRUS prostate glands showing the overlap of the transformed source (MR) and target (TRUS) point-sets. The green shape illustrates the target point-set, while the red shape illustrates the transformed source point-set. The first column shows the source and target after center-alignment. The remaining columns show registrations from various methods; from left to right: FPT-Chamfer, FPT-Chamfer trained with Scenario 1 image slices, FPT-Chamfer trained with Scenario 2 image slices, and FPT-Chamfer trained with Scenario 3 image slices. Figure from [120].

draw further conclusions.

Though there is a measurable difference in the mean TRE, $D_C$, and $D_H$ obtained between the full volumetric registrations and partial registrations for FPT-Chamfer, it is notable that these variations provide little qualitative difference, as seen in Figures 3.14 and 3.15; only sub-millimeter differences in quantitative performance were observed between each of the three clinical scenarios. This highly comparable performance demonstrates FPT’s flexibility and generalizability between different input data and illustrates that the network can adapt to multiple varied distributions and availabilities of input data and still learn to predict a desirable registration.

Intensity-based methods for multimodal image registration are also able to utilize information from the entire prostate gland, typically providing a qualitatively and quantitatively good intraprostatic deformation. To emulate this, volumetric point-sets were utilized, as this allows the network to learn intraprostatic deformation instead
3.6. Prostate MR-TRUS Registration

Figure 3.15: Example image slices illustrating registration results from three different cases, shown in the transverse plane. Each image shows the original TRUS image with the transformed source (MR) contours (red) superimposed onto the target (TRUS) contours (green). Columns indicate registrations from FPT-Chamfer for registrations with; from left to right; full volumes, Scenario 1, Scenario 2, and Scenario 3. Figure from [120].

Recently, intensity-based deep learning methods have achieved reported TREs below 5 mm for the MR-TRUS registration application explored in this work [114, 137, 169, 206]. The TREs obtained for this application in this work fall within the expected range previously defined clinically significant thresholds of 2.97 mm [208] and 5 mm [200] found in the literature. However, it is difficult to make direct comparisons between these results and others due to variations in the quality of data (for example, due to different clinical setups, image acquisition protocols, and user experience) and validation methods. In particular, the number and spatial distribution of landmarks used to estimate TRE is likely to have a significant impact on the numerical error. In the dataset used, the landmarks used to calculate TRE, such as the apex and base of the prostate, are located on the surface or towards the periphery of
the prostate gland (unlike the urethra, for instance). Furthermore, the aforementioned works do not consider the practical scenario of primary interest in this work, where only partial data are available due to a limited number of image slices. An important finding of this study is the minimal impact of using partial point data on accuracy when using the FPT method compared to other methods tested. Without extensive validation, it is unclear if the performance of intensity-based methods and/or other forms of representations, such as binary masks, would also be minimally impacted by this reduction of data. As such, the practical effects of partial data when applied to existing registration methods and frameworks merit a thorough validation and
3.7 Conclusion

This chapter has introduced FPT, a novel approach to point-set registration using deep neural networks which learns the displacement field required to produce individual point displacements. Through evaluation with synthetic non-medical data from computer vision benchmarking datasets, US-based spine atlas reconstructions, and in a challenging real-world multimodal image registration task with MR and TRUS images, FPT was found to be robust to deformation, noise, and the partial availability of data; demonstrating multiple real-world use cases and clinical applications. Through the evaluation of atlas-based registration to US-based spine reconstructions, FPT has been shown to be generalizable to geometries external to its training data domain. Furthermore, this work demonstrates that partially available data, generated from automatically segmented MR and TRUS images, may be used to enable continual real-time MR-TRUS image registration during prostate biopsies. In other medical imaging problems where training data may be limited, FPT’s generalizability may be of interest, given its ability to rapidly register point-sets extracted from imaging acquired at different times or from different modalities. This demonstrates FPT’s
utility as a generally-applicable method for learning-based non-rigid registration, representing significant progress for non-iterative, non-rigid point-set registration. As a registration method that non-iteratively performs non-rigid registration without needing established point correspondence, FPT also represents significant progress towards a generally applicable method for learning-based non-rigid registration in medical imaging.
Part II

Intensity-based Multimodal Registration
Chapter 4

Meta-Learning Initializations for Interactive Registration

This chapter is based on the work “Meta-learning initializations for interactive medical image registration”, published in IEEE Transactions on Medical Imaging [207].

4.1 Introduction

4.1.1 Novel Training Paradigms for Learning-Based Medical Image Registration

Beyond the use of methodologies from “classical” iterative registration algorithms, learning-based methods have recently been proposed for medical image registration in ways that enhance performance, relative to non-learning-based methods, while generalizing well to unseen images. All this, while maintaining their rapid inference and substantially reduced computational requirements at test-time. While learning-based registration methods have been proposed to use different network architectures, such as convolutional neural networks [114, 209] and vision transformers [210], they have additionally been proposed with different training objectives and strategies, such as generative adversarial networks [137, 167], supervised [114, 137], unsupervised [119, 120, 209, 211] or reinforcement learning [114, 169, 212]. Furthermore, depending on the available data and the application, parametric transformations may be used which are spline-based [211], or with and without further
desirable constraints, such as diffeomorphic [213] or biomechanical [118] constraints. Since the registration has been formulated as a machine learning problem for directly predicting transformation between a given image pair, methodologies such as semi-supervised learning [214], few-shot- and meta-learning [215], unsupervised contrastive learning [216], inference-time augmentation [217], and amortized hyperparameter learning [218] have also been trialed to improve the data efficiency and generalizability of the trained registration networks.

4.1.2 Interactive Machine Learning

Learning-based methods often achieve state-of-the-art performance in medical image registration [107]. However, this performance is often demonstrated on well-curated datasets which closely match the distribution of the training data. As such, reported performance may not consistently demonstrate the generalization, robustness and accuracy required for clinical use [108]. Often, these differences in performance are due to the challenges associated with the deployment of medical image analysis methods, such as inconsistent image quality, varied imaging protocols, and interpatient variation [219]. In many cases, issues may be efficiently and effectively identified or corrected by an experienced human observer. When errors may be corrected via user interaction, their integration into a conventional machine learning framework may assist in predicting more accurate solutions [220–223]. This integration is referred to as interactive machine learning (IML); the design and implementation of learning-based methods and their accompanying interfaces which integrate user interactions to guide prediction [224] (Figure 4.1). In IML, user interactions may take various forms, and may or may not require additional steps of gradient descent to update the trained model to yield a more accurate prediction. Potential interactions and their use in refining trained models are further discussed in Sections 4.3.2 and 4.4.3, respectively.

Performance in challenging real-world interventional or surgical registration tasks may be assisted by IML. Though it may be undesirable to couple interactions with real-time processes, interactions may compensate for deficiencies in the method that are difficult or even infeasible to account for during training. However, inter-
4.1. Introduction

Figure 4.1: A schematic of conventional machine learning and interactive machine learning approaches at inference. In the conventional approach (top), previously unseen test data is passed through a trained model to produce a prediction. In contrast, the interactive machine learning approach (bottom) leverages user interactions to provide feedback to the model, based on the intermediate prediction, to obtain a more accurate prediction.

action may increase computational complexity and cognitive effort and needs to be carefully weighed against its performance gain. Therefore, such interactions must be simple and feasible to acquire for training, for example, by using simulated interactions [220–223]. Recently, IML-based methods for medical image analysis have predominantly performed error correction with user interaction for image segmentation [220–223]. To be viable for practical application, most existing methods used simple annotations as user interactions. UI-Net [220] and iUNet [223] integrate ‘scribbles’ as inputs to indicate areas that should (i.e. False Negatives) or should not (i.e. False Positives) be considered by the segmentation method when refining the segmentation, with simulated and user-provided scribbles for training and inference, respectively. BIFSeg combines an initial user-defined bounding box as an input to guide the initial prediction, and scribble-based interactions during inference [221]. DeepIGeoS takes a different approach that uses two networks; one for initial segmentation and another for refinement [222]. Though there is a lack of existing IML-based methods for medical image registration in the literature, interaction has previously been utilized in non-learning-based methods for interactive registration. Such methods focus on the interactive selection of anatomical landmarks [225] or spatially
4.1. Introduction

tracked intra-operative surgical instruments [226], to improve the alignment of the initially fused patient images. Though these aforementioned registration methods are not learning-based, both are still widely used and are considered to be among the gold-standard for interactive medical image analysis methods - demonstrating the practicality of coupling interactions with learning-based registration methods.

4.1.3 Gradient-Based Meta-Learning

A neural network may be pre-trained with a large set of natural images (e.g. ImageNet [227]) to obtain an initialization. From this initialization, a new network may then be trained, or ‘fine-tuned’, on a similar domain to reduce training time or the amount of data required. This approach has been shown to deliver, at worst, comparable performance to training networks from scratch [228].

Meta-learning [229, 230] can be used to formalize the pre-training-then-fine-tuning intuition for improving fine-tuning capabilities by iteratively learning how to improve future performance on a distribution of related tasks over multiple learning episodes. In particular, approaches categorized as ‘gradient-based’ meta-learning, such as Model-Agnostic Meta-Learning (MAML) [231] and Reptile [232], learn easily adaptable initializations from the gradients observed during a learning episode. By allowing for direct optimization of performance to new labeled data from “future” unseen tasks during the adaptation (or fine-tuning) stage, meta-learning contrasts conventional machine learning methods which optimize for generalized performance, without the use of any adaptation stage.

These aforementioned gradient-based meta-learning algorithms are simple, learn quickly, and generalize well at test time with limited examples, as evidenced by their successful application in the medical imaging domain [215, 233–236]. Park et al. [215] utilize Reptile to learn domain-agnostic initializations for medical image registration which adapt rapidly to unseen domains with limited data. Liu et al. [233] introduce shape-aware meta-learning for domain generalization by optimizing the compactness and smoothness of the segmentations under simulated domain shift to allow for adaptation to images from new datasets. Similarly, Khandelwal and Yushkevich [234] and Khadga et al. [235] utilize MAML as a means to learn a segmentation
initialization that may be quickly adapted to other datasets. Additionally, novel gradient-based meta-learning algorithms have been shown to outperform MAML and Reptile for rapid domain adaptation in medical image segmentation tasks [236]. It is of note that such efforts utilize generalized domain-agnostic information to enable rapid training, and improved performance in unseen domains, not to improve domain-specific performance.

Contrarily to these aforementioned examples, in this work, a focus is placed on improving performance within the domain of a specific task, as the author was not able to identify any existing domain-specific meta-learning-based imaging registration methods in the literature. The use of an existing gradient-based meta-learning algorithm, Reptile, makes sense - given its simplicity in incorporating new-task data, efficiency in adaptation and the afore-summarized effectiveness in various computer vision and medical imaging applications. These are important features to adopt when formulating an interactive registration framework, such as the framework described in the following sections. Other meta-learning methodologies, such as those based on bi-level optimization [231, 237] and reinforcement learning [134], should be tested in future development. It is interesting to note that several existing interactive segmentation algorithms, described in Section 4.1.2, may also be considered as instances of meta-learning algorithms.

4.2 Contributions

This chapter describes a framework to meta-learn network parameter initializations for interactive image registration. In practice, this framework consists of three components:

1. A learning-based medical image registration algorithm,

2. A form of user interaction that refines predictions at inference and is easily simulated during training, and

3. A gradient-based meta-learning protocol that learns a rapidly adaptable network initialization.
4.3. Interactive Medical Image Registration with Meta-Learning

This framework is unlike other existing works as it seeks to utilize and combine the error-correcting benefits of user interaction with the rapid adaptation capabilities of gradient-based meta-learning to optimize registration performance within a single target domain, by considering individual patients as separate tasks.

To investigate the application of such a framework to clinical data, this work seeks to register 3D MR volumes to a series of interactively-acquired sparse 2D TRUS images for use in targeted prostate biopsy guidance. This exemplar application illustrates a clinical scenario in which real-time, interventional imaging, such as TRUS, is acquired interactively to iteratively refine the registration throughout a single acquisition of interventional imaging modality as it traverses the target anatomy. This work aims to compare the accuracy of the proposed interactive registration method with alternative learning-based methods.

First, a detailed description of the framework for meta-learning an initialization for interactive medical image registration and describe how such an interactive framework may enable a wide range of practically useful applications. Subsequently, the registration, interaction, and meta-learning strategy are introduced and described for the exemplar clinical application. In addition, a novel implementation of a learning-based volume-to-sparse-slices registration algorithm is also introduced to align MR to TRUS images in this application. Lastly, rigorous analysis and validation experiments are presented which compare the proposed methodology to different learning-based methods for the prostate MR-TRUS image registration application, including validation on multiple variations to the meta-learning parameters to assess their effects on the interactive registration process.

4.3 Interactive Medical Image Registration with Meta-Learning

The proposed framework for interactive medical image registration with meta-learning is described below by each of its three components, as previously summarized in Section 4.2. In this section, the general motivation, rationale and formulation for each component are proposed, leading to the specific algorithm proposed in
Section 4.4.

4.3.1 Learning-based Image Registration

Learning-based registration may be categorized from an imaging application perspective, network inputs may be unimodal, multimodal, inter-patient, or intra-patient - with each image bearing its own dimensionality [107], which in turn requires different types of loss functions, including those based on unsupervised image similarity [211], weakly-supervised label similarity [114], or some combination of the two [209]. Moreover, each image pair may encompass any number of anatomical sites of clinical interest, requiring a registration method to utilize different deformation models, commonly; rigid, affine, or deformable [107].

Given \( N \) pairs of training source and target images, \( \{x_n^{\text{source}}\} \) and \( \{x_n^{\text{target}}\} \), and accompanying source and target labels, \( \{l_n^{\text{source}}\} \) and \( \{l_n^{\text{target}}\} \), respectively, where \( n = 1, \ldots, N \), existing approaches predict the voxel correspondence or transformation \( u_n^{\phi} = f^{\phi}(x_n^{\text{source}}, x_n^{\text{target}}) \) using a registration network \( f^{\phi} \) with network parameters or weights \( \phi \). The training goal thus is minimizing an image and/or label loss function \( L_{\text{sim}} \) over \( N \) training pairs, to obtain the optimal \( \hat{\phi} \):

\[
\hat{\phi} = \underset{\phi}{\text{argmin}} \sum_{n=1}^{N} \left[ L_{\text{sim}}(\phi) + \alpha^{\text{def}} L_{\text{def}}(\phi) \right],
\]

where \( L_{\text{def}} \left( \phi \mid x_n^{\text{source}}, x_n^{\text{target}} \right) = L_{\text{def}} \left( f^{\phi}(x_n^{\text{source}}, x_n^{\text{target}}) \right) \) provides regularization on the smoothness of the deformation \( u_n^{\phi} \) weighted by a hyperparameter \( \alpha^{\text{def}} \). In general, the similarity-based loss can further combine a negative unsupervised image similarity function \( L_{\text{sim}}^{\text{image}} \left( x_n^{\text{source}}(u_n^{\phi}), x_n^{\text{target}} \right) \), between the transformation-warped images \( x_n^{\text{source}}(u_n^{\phi}) \) and the target images \( x_n^{\text{target}} \), and a negative weak-supervision loss based on label similarity \( L_{\text{sim}}^{\text{label}} \left( l_n^{\text{source}}(u_n^{\phi}), l_n^{\text{target}} \right) \), between the warped source labels \( l_n^{\text{source}}(u_n^{\phi}) \) and the target labels \( l_n^{\text{target}} \):

\[
L_{\text{sim}}(\phi \mid x_n^{\text{source}}, x_n^{\text{target}}, l_n^{\text{source}}, l_n^{\text{target}}) = \alpha^{\text{image}} L_{\text{sim}}^{\text{image}} \left( x_n^{\text{source}}(u_n^{\phi}), x_n^{\text{target}} \right) + \alpha^{\text{label}} L_{\text{sim}}^{\text{label}} \left( l_n^{\text{source}}(u_n^{\phi}), l_n^{\text{target}} \right).
\]
where hyperparameters $\alpha_{\text{image}}$ and $\alpha_{\text{label}}$ can be set to zero to represent the previously proposed weak supervision and unsupervised algorithms, respectively. This general formulation includes both image- and label-based losses for the learning-based registration methods to permit the formulation of a general interactive registration, as described in the remainder of this chapter. However, other methods such as those based on adversarial learning and reinforcement learning may need further adaptation to incorporate interaction.

### 4.3.2 Interaction for Image Registration

In general, the performance improvement seen in other interactive applications, such as the above-discussed interactive segmentation [233–236], may be expected from interactive registration. Other benefits, such as those related to expandability, and owing to the human-in-the-loop of machine learning models for registration applications are also important but are considered out of the scope of this work.

To adapt existing learning-based registration methods to accept interactions, potential interactions which can be learned during training must first be defined. In this work, interaction is considered to be any action, taken by one entity (i.e. the user or an automated computer algorithm) that has a reciprocated action taken by the other entity as a result of the initial action. Notably, each interaction may be formulated as a collection of interactions; either as some combination of the same or of different types of interaction. Depending on application-specific needs, the combination of multiple interactions may best provide additional information for improving registration or, equivalently, error correction. This interactive process should be learnable during model training and feasible at test-time. Below, several instances of possible registration-based interactions are described across two categories of actions; namely, computer-to-user and user-to-computer. Both categories are formulated as the addition of new data, using a single form of interaction to facilitate interactive learning-based algorithms.

When a computer algorithm takes an action (e.g. makes a predictive registration, in this case), the user provides a reciprocated interaction for error correction. Following this correction, a new prediction may be immediately made. For example,
image re-acquisition or annotation within areas that are not well aligned. Image re-acquisition can occur on a local (i.e. one, or a few images) or global scale (i.e. the entire image volume) when, for example, image quality is poor, or there has been patient motion. Image re-acquisition may be especially pertinent when using real-time imaging modalities, such as US, given that image quality may be operator dependent, and that images may be rapidly re-acquired. This computer-to-user interaction enables user-defined annotations to indicate regions in the registration that are poorly aligned.

This work proposes that both image reacquisition and annotation may be formulated generally as additional labeled data, where the quantity and availability of labels and images may vary depending on different applications. In practice, to learn an interactive registration model, the utility of each type of interaction may be highly application-specific. The determination of the most practical use-cases for each interaction is considered out of the scope of this thesis, though in this chapter, a thorough description and evaluation of the addition of new data interactions in an TRUS-guided prostate biopsy application is provided to illustrate the use of an interactive registration learning framework.

4.3.3 Meta-Learning Interactive Initializations

Learning interactive medical image registration through a meta-learning protocol allows the model to learn an easily adaptable initialization from which rapid and task-specific fine-tuning may occur, instead of simply fine-tuning or adapting the input on a conventionally-trained model. In this work, it is proposed that during adaptation the model accepts the aforementioned newly labeled data and provides a reciprocated action. Following this addition of data, a new prediction may be immediately made, or some task-specific fine-tuning may first occur before giving a refined prediction, thus improving the registration.

As discussed in Section 4.3.2, such newly labeled data may include image acquisition, annotation, or indicate areas that should be aligned or have already been well-aligned. As such, the use of a meta-learning protocol permits the interactive model to be trained in such a way that it learns how to utilize this new data to best
predict a better registration. This may be achieved by fine-tuning the same set of network parameters in an outer-loop of the meta-learnin for learning the adaptation ability across different interaction tasks, together with the inner-loop that optimizes the registration network for the given task. Details of a general implementation based on gradient-based meta-learning are proposed in Section 4.4.2.

4.4 Methods

4.4.1 Images and Annotations as Interaction

Following the discussion in Section 4.3.2, possible pairs of interactions which are “sampled” from the source and target images are denoted as \{i_{mn}^{\text{source}}\} and \{i_{mn}^{\text{target}}\}, from \(n\) training data. Each \(n^{th}\) pair is also associated with \(M_n\), the number of possible interactions that are possible on image pair \(n, m = 1, \ldots, M_n\), respectively.

Without loss of generality, it is proposed to represent these time-agnostic interactions as sets of interactively obtained images (\{\mathbf{x}_{mn}^{\text{source}}\} and \{\mathbf{x}_{mn}^{\text{target}}\}) and annotations which are generally in the form of segmentation labels (\{\mathbf{\ell}_{mn}^{\text{source}}\} and \{\mathbf{\ell}_{mn}^{\text{target}}\}), i.e.,

\[i_{mn}^{\text{source}} = [(\mathbf{x}_{mn}^{\text{source}})^{T}, (\mathbf{\ell}_{mn}^{\text{source}})^{T}]^{T}\]  

and  

\[i_{mn}^{\text{target}} = [(\mathbf{x}_{mn}^{\text{target}})^{T}, (\mathbf{\ell}_{mn}^{\text{target}})^{T}]^{T}\].

For notational brevity, both images and annotations can include the previously available annotated data, for individual subject, therefore the interactions \{i_{mn}^{\text{source}}\} and \{i_{mn}^{\text{target}}\} are interchangeably used with interaction-updated source and target, respectively. A sequence of interactions may benefit from explicit sequential modeling; however, this is considered out of scope of this work, where only a few steps of interaction are considered feasible in the application of interest.

This formulation does not distinguish between registrations which may have different initial image and annotation data from one without such initial registration, as they can be consistently represented by both the non-interactive registration formulation, described in Section 4.3.1, and the interactive adaptation, described in Section 4.4.2.

It is of note that not all the interactive image or annotation data need to be available or varying for a given interaction. Below, a sample of scenarios that demonstrate the versatility of interactive registration are described. Additionally,
active learning methodologies [108] may appear similar in nature and may be able to utilize similar scenarios for interactive learning in practice. This application is developed and validated with respect to Scenario 4, a special case of Scenario 3. Though not tested, other scenarios are included for discussion purposes.

1. A user may iteratively add successive annotations to an image to improve the registration between two images over multiple interactions, i.e. variable labels \( \ell_{\text{source}}^{m=a, n} \neq \ell_{\text{source}}^{m=b, n} \) and \( \ell_{\text{target}}^{m=a, n} \neq \ell_{\text{target}}^{m=b, n} \) with fixed images \( x_{\text{source}}^{m=a, n} = x_{\text{source}}^{m=b, n} \) and \( x_{\text{target}}^{m=a, n} = x_{\text{target}}^{m=b, n} \), when \( a \neq b \).

2. An unsupervised learning algorithm, without initial labels \( \ell_{\text{source}}^{m=0, n} \) and \( \ell_{\text{target}}^{m=0, n} \), may benefit from further interactively-defined annotations to improve the alignment, which requires the simulation of \( \ell_{\text{source}}^{m>0, n} \) and \( \ell_{\text{target}}^{m>0, n} \) during training to mimic such a scenario.

3. An image-guidance application may have a fixed pre-operative image \( x_{\text{source}}^{\text{pre-op}} \), but with potential interactions to add new intra-operative images, i.e. \( x_{\text{source}}^{m=a, n} \neq x_{\text{source}}^{m=b, n} \) and \( x_{\text{target}}^{m=a, n} = x_{\text{target}}^{m=b, n} \), when \( a \neq b \). This same application may use additional interactively-defined annotations on both the pre-operative and intra-operative images, i.e. variable source and target labels, as in Scenario 1.

4. A TRUS-guided prostate cancer application, such as that used in this work and further described in Section 4.4.2, may be similar to Scenario 3, but does not require the use of, or generation of, additional annotations on the source images, whilst the additional annotations on the target images may be acquired automatically using an independent prostate TRUS segmentation network, i.e. using the generation of labelled intra-operative TRUS images as the interaction, \( x_{\text{source}}^{m=a, n} = x_{\text{source}}^{m=b, n} \), \( x_{\text{source}}^{m=a, n} = x_{\text{source}}^{m=b, n} \), \( x_{\text{target}}^{m=a, n} = x_{\text{target}}^{m=b, n} \), \( \ell_{\text{source}}^{m=a, n} \neq \ell_{\text{source}}^{m=b, n} \) and \( \ell_{\text{target}}^{m=a, n} \neq \ell_{\text{target}}^{m=b, n} \), when \( a \neq b \).

4.4.2 Meta-Learning for Interactive Registration

As the interaction data \( \{\ell_{mn}\} \) and \( \{\ell_{mn}'\} \) are defined as images and annotations - in Section 4.4.1 - \( \{x_{mn}\} \), \( \{x_{mn}'\} \), \( \{\ell_{mn}'\} \) and \( \{\ell_{mn}'\} \), they are consistent
with the data used in the non-interactive registration formulation - in Section 4.3.2 - \( \{ x_n^{\text{source}} \}, \{ x_n^{\text{target}} \}, \{ I_n^{\text{source}} \} \) and \( \{ I_n^{\text{target}} \} \). The proposed formulation for training an interactive registration network \( f_{\hat{\phi}} \) results from adapting the optimization given in Eq. (4.1) to a bi-level optimization \([230, 238]\), therefore learning the interactive image registration becomes a meta-learning problem:

\[
\hat{\phi} = \arg\min_{\phi} \sum_{n=1}^{N} \sum_{m=1}^{M_n} \left[ L_{\text{sim}}^* (\phi^* (\phi)) + \alpha_{\text{def}} L_{\text{def}}^* (\phi^* (\phi)) \right], \tag{4.3}
\]

s.t. \( \phi^* = \arg\min_{\phi} \sum_{n=1}^{N} \sum_{m=1}^{M_n} \left[ L_{\text{sim}}^* (\phi) + \alpha_{\text{def}} L_{\text{def}}^* (\phi) \right], \tag{4.4}
\]

where the similarity term \( L_{\text{sim}}^* \) is obtained from the substitution of interaction data into Eq. (4.2):

\[
L_{\text{sim}}^* = L_{\text{sim}} (\phi \mid x_{mn}^{\text{source}}, x_{mn}^{\text{target}}, \ell_{mn}^{\text{source}}, \ell_{mn}^{\text{target}}), \tag{4.5}
\]

similarly, \( L_{\text{def}}^* = L_{\text{def}} (\phi \mid x_{mn}^{\text{source}}, x_{mn}^{\text{target}}) \). \( L_{\text{sim}}^* (\phi^* (\phi)) \) and \( L_{\text{def}}^* (\phi^* (\phi)) \) denote the initialized functions of \( \phi \), by optimized \( \phi^* \) at the inner-level. \( \phi^* \) is hereinafter used for brevity.

It is noteworthy that, unlike the training defined in Eq. (4.1) which minimizes the expected loss over the \( N \) pairs of training images, the task-specific inner-level Eq. (4.4) aims to minimize the expected loss over the \( M_n \) samples of interactions. At the outer-level, Eq. (4.3), different \( N \) pairs of images and annotation are usually sampled to learn the optimal network parameters, such that, at inference, the network \( f_{\hat{\phi}} \) can be adapted to new pairs of interactions \( \{ I_{m, \text{test}}^{\text{source}} \} \) and \( \{ I_{m, \text{test}}^{\text{target}} \} \), where \( m = 1, \ldots, M_{\text{test}} \) and be generalized to this new test task, i.e. the training meta-tasks are defined as the \( N \) different cases that need registration, rather than \( M_n \) steps of interactions.

Such a meta-learning framework learns an initialization of network parameters \( \hat{\phi} \) which enables data-efficient adaptation to a new task at inference. The efficient adaptation means that registering a new pair of images \( x_{\text{test}}^{\text{source}} \) and \( x_{\text{test}}^{\text{target}} \) may only require a few \( M_{\text{test}} \) steps of interaction, often constrained by human effort and
time-critical applications.

4.4.3 Gradient-Based Meta-Learning Algorithms for Network Initialization

Gradient-based meta-learning algorithms, discussed in Section 4.1.3, are applicable for training the proposed interactive registration networks and are comprised of the meta-learning and the meta-test phases. At the start of meta-training, the registration model is initialized with random weights. Then, during each iteration of the outer-level loop, one task \((i_{mn}^{\text{source}}, i_{mn}^{\text{target}})_n\) is randomly sampled from the task set \(\{(i_{mn}^{\text{source}}, i_{mn}^{\text{target}})_n = 1, ..., N\}\) that contains all available tasks, with a set of \(k\) interactions \(\{(i_{mn}^{\text{source}}, i_{mn}^{\text{target}})_n, m = 1, ..., k\}\) randomly sampled from this given task, to form an episode (Figure 4.2). Each sampled task corresponds to a task-specific loss in Eq. (4.4) using only data from this task, the meta-learning task is defined as a pair of source and target images with their associated source and target annotations to be registered. During each episode, ‘task-level learning’ occurs using Stochastic Gradient Descent (SGD), or a similar variant (e.g. Adam [185]), for \(k\) SGD steps, the task-specific gradient \(g^m_n(\phi)\) can be computed to update the network weights \(\phi\):

\[
\phi^*_m \leftarrow \phi - \beta_{\text{task}} \cdot g^m_n(\phi), \tag{4.6}
\]

where

\[
g^m_n(\phi) = \frac{\partial}{\partial \phi} \left[ \mathcal{L}_{\text{sim}}^* (\phi) + \alpha_{\text{def}}^* \mathcal{L}_{\text{def}}^* (\phi) \right], \tag{4.7}
\]

and \(\beta_{\text{task}}\) is the learning rate. After an episode of \(k\) steps, a cross-task gradient \(g_n(\phi^*)\) is used to update the network weights at the outer-level loop, corresponding to Eq. (4.3):

\[
\phi_n \leftarrow \phi - \beta_{\text{meta}} \cdot g_n(\phi^*), \tag{4.8}
\]

where

\[
g_n(\phi^*) = \frac{\partial}{\partial \phi} \left[ \mathcal{L}_{\text{sim}}^* (\phi) + \alpha_{\text{def}}^* \mathcal{L}_{\text{def}}^* (\phi) \right] (\phi^*), \tag{4.9}
\]

and \(\beta_{\text{meta}}\) is the meta-learning rate. With gradient-based meta-learning methods,
4.4. Methods

Figure 4.2: Diagram illustrating one episode of task-level learning wherein a task $(i_{mn}^{source}, i_{mn}^{target})_n$ is sampled. For each interaction of the $k$ interactions in the sampled task, the image pair $\{x_{mn}^{source}\}, \{x_{mn}^{target}\}$ is coupled with an associated annotation $\{\ell_{mn}^{source}\}, \{\ell_{mn}^{target}\}$. Figure from [207].

such as MAML [231], the cross-task meta-gradient $g_n(\phi^*)$ is computed directly to obtain the Jacobian for updating parameters, at the inner-loop-optimized weight values $\phi^*$. However, estimating the Jacobian involves computationally problematic second derivates; First-Order MAML [231] and Reptile [232] have been proposed to approximate this meta update step, and this work adapts such approximations to train the proposed interactive registration network.

In the meta-test phase, parameters $\tilde{\phi}$ are adapted to the test task through a few-shot learning process. During meta-testing, a few interactions $\{i_{m, test}^{source}\}$ and $\{i_{m, test}^{target}\}$ are acquired directly from the test task to compute a few steps of test-task-specific gradients, to update the model using Eq. (4.6), before it predicts the transformation using the pair of images $x_{test}^{source}$ and $x_{test}^{target}$, as illustrated in Figure 4.3.
4.4. Methods

Figure 4.3: Schematic representation of a framework for interactive medical image registration with meta-learning. A learning-based registration method is trained over multiple episodes during the meta-training phase (left) to learn an initialization for adaptation at inference. During each task-level learning episode, a task is sampled; comprising a set of image and annotation pairs. Following each episode, the meta-update (red arrow) updates the learning-based registration method based on the direction (dashed line) of the $k$ task-level learning gradients (white arrows) and continues onward from the previously learned gradients (blue arrows). Once training is complete, the registration method is optimized in the meta-test phases (right). Here, few-shot learning on the target data, coupled with user-defined interactions, yields a fine-tuned registration model. Figure from [207].

4.4.4 Exemplar Clinical Application

In this section, the application of the proposed meta-learning framework for interactive registration is presented for a real-world clinical application. The selected methods and implementations are described for prostate MR-TRUS image registration, where only sparse TRUS images are available, within the context of a meta-learning framework for interactive weakly-supervised multimodal image registration.

Prostate MR-TRUS image registration is a method for leveraging MR imaging to aid in performing tumor-targeted needle biopsy [195, 239–245] and minimally-invasive treatments [196, 246] when it is suspected or confirmed that a patient has clinically significant prostate cancer. Image registration techniques allow the presentation of MR-visible information such as tumor size and location. This permits needles and other instruments to be placed advantageously to target specific tissue
regions within the patient’s prostate. Often, this is displayed as a visual overlay, where the MR-derived lesion and tumor information is superimposed on the TRUS images as a composite image or is presented alongside the TRUS image. Displaying these images simultaneously in an effective and useful manner demands an accurate registration.

4.4.5 Volume-to-Sparse Weakly Supervised Multimodal Image Registration

A weakly-supervised training methodology is a special case of the general formulation, as discussed in Section 4.3.1, in turn, training an interactive registration network with a label-driven loss can be considered as a meta-learning problem described in Eq. (4.3) and Eq. (4.4), with $\alpha_{image} = 0$, without using explicit intensity-based similarity measures which have been considered less effective [114]. Furthermore, accommodating sparse US images that are readily available as part of the interactions in this application, a volume-to-sparse registration algorithm is first developed, where the training target images being a set of TRUS slices $\{x_{mn}^{target}\}$ and annotations of anatomical structures identified on these slices $\{I_{mn}^{target}\}$, with source MR images $\{x_{mn}^{source}\}$ and the corresponding MR annotations $\{I_{mn}^{source}\}$. These annotations can contain multiple types of anatomical structures, with similar applicability discussed in the weak supervision algorithm [114]. This notation is however omitted in this work, for brevity. A discussion of the detailed representation of the counterpart interactive data is given in Section 4.4.6, and the need for the availability of TRUS slice location information is further summarized in Section 4.7.

In this implementation, a recent method for weakly-supervised image registration called LocalNet [114] is utilized. LocalNet has an encoder-decoder structure comprised of four down-sampling blocks and four up-sampling blocks and can predict a Dense Displacement Field (DDF) that may be summed over multiple resolutions. LocalNet is similar to the UNet [110] architecture found in VoxelMorph [209] — often used for unsupervised and weakly-supervised image registration. Compared to UNet, LocalNet has a smaller memory requirement and is more densely connected, featuring multiple types of residual shortcuts and summation-based skip layers to
allow the model to benefit from deeper supervision [114].

4.4.6 Interactive Acquisition of Labelled TRUS Images

Conventionally, when registering MR and TRUS images to guide tumor-targeted needle biopsies, the acquisition of a complete 3D TRUS image volume is required. However, this presents some unique challenges, given the inherent difficulty in acquiring a dense set of images from which the volume may be reconstructed. In cases where this initial acquisition is not sufficient for registration, an entirely new image volume is typically acquired. Therefore, this work proposes a new paradigm for MR-TRUS registration whereby a volume-to-sparse registration continually re-occurs throughout the acquisition, preventing the need to subsequently reacquire any images or image volumes. As such, to augment a volume-to-sparse registration framework with continuous registration throughout the acquisition, the addition of new data, in the form of TRUS images and their accompanying automatically acquired prostate gland segmentation, using TRUS segmentation methods such as [197], as an interaction is proposed. As such, at inference, this continuous stream of interactively acquired data provides the model with additional information, context, and the potential to compute a constantly up-to-date registration, possibly yielding more accurate registration results.

In this work, the interaction stems from the continual acquisition of frames in a single sweep of the TRUS probe. Therefore, during few-shot learning, each new frame is incorporated into the image volume as part of the input to the model. This requires knowledge of the spatial relationship between each frame, so that the new frame may be inserted into the correct location within the TRUS volume. To provide initial spatial information for the network, the first interaction comprises two frames, and subsequent interactions require only one new frame. By gathering images in sequence, the ability to fine-tune a registration model based on patient-specific data becomes feasible. This may prevent the need to capture a complete 3D TRUS volume as several sparse images may provide comparable registration accuracy. Given the current clinical workflow for tumor-targeted needle biopsies, the selected interaction may not introduce any additional time delay or modify existing protocols.
To simulate this interaction during training, one pair of target interactions \( \mathbf{y}_{mn}^{\text{target}} \) is selected by randomly selecting a series of TRUS images in a clinically feasible manner, whilst the target “interaction” is the fixed MR images and their annotation \( \mathbf{y}_{mn}^{\text{source}} \), as described in Scenario 4 in Section 4.4.1. The label pair \( \mathbf{y}_{mn}^{\text{source}} \) and \( \mathbf{y}_{mn}^{\text{target}} \) may define either the prostate boundary, the apex and base of the prostate, or any other patient-specific landmarks; such as zonal structure boundaries, water-filled cysts, and calcifications [146]. The binary mask is randomly generated to include some number of randomly selected frames \( F \), where \( F \in \mathbb{N} : F \in [F_{\min}, F_{\max}] \), which defines the image slices within the TRUS volume \( \mathbf{x}_{mn}^{\text{target}} \). Once generated, this binary mask is used to mask out sections of the input image \( \mathbf{x}_{mn}^{\text{target}} \), and corresponding label \( \ell_{mn}^{\text{target}} \), leaving only TRUS slices and corresponding labels which are obtained from the simulated acquisition.

### 4.4.7 Meta-Learning an Initialization with Reptile

A visual summary of the below described meta-learning phases for the presented application is shown in Figure 4.4.

#### 4.4.7.1 Meta-Training Phase

Reptile [232] is adopted as the gradient-based meta-learning strategy for the proposed interactive registration framework. Reptile provides a computationally efficient optimization of the gradient-based update procedure to approximate Eq. (4.8) and Eq. (4.9):

\[
\phi_n \leftarrow \phi - \beta^{\text{meta}} \cdot \sum_{m=1}^{k} (\phi - \phi_m^*),
\]

(4.10)

where \( \phi_m^* \) can be estimated using Eq. (4.6). In this work, the Adam optimizer [185] is used in the meta-optimization.

It is interesting to note that, given that the complete prostate (or other patient-specific landmarks) labels are available, a stronger form of supervision is employed in this work to compute the losses during the meta-training phase, such that the similarity of the entire label is computed instead of a partial similarity on the sparse labels. This allows the initialization to be learned from complete data, facilitating
Figure 4.4: Schematic representation of the proposed framework for interactive medical image registration with meta-learning as applied to a weakly-supervised volume-to-sparse prostate MR-TRUS registration problem. The weakly-supervised learner is trained over multiple episodes during the meta-training phase (left) to learn an initialization for adaptation at inference. During each task-level learning episode, a task is sampled; comprising a set of images, labels, and some number of frames $F$. After each episode, the meta-update step updates the learner using the Reptile algorithm based on the task-level learning gradients. Once training is complete, the learner is optimized during the meta-test phases (right). Here, interactively-acquired, patient-specific, data is coupled with few-shot learning to yield a fine-tuned registration model, in real-time, as the TRUS image acquisition occurs. Figure from [207].

guidance by a loss that incorporates complete segmentation labels. This illustrates how the interactive labels and images may be used in computing training losses may differ from those seen in the meta-test phase in order to better guide learning, benefitting the adaptation capabilities during meta-testing.

4.4.7.2 Meta-Test Phase

During the meta-test phase, for evaluation, few-shot learning is completed with $F$ gradient updates on interactions $\mathbf{x}_{mn}^{\text{target}}$ and $\ell_{mn}^{\text{target}}$ sampled from the test task. This fine-tunes the registration model to obtain adapted parameters $\phi'$ which can perform accurate registrations on the test patient. Unlike the random generation of interactions during the meta-training phase, $\mathbf{x}_{mn}^{\text{target}}$ and $\ell_{mn}^{\text{target}}$ define a continuous TRUS acquisition. Therefore, the first few-shot learning gradient update contains
$F_{\text{min}}$ images and each subsequent update adds an image, until the final update, comprised of $F_{\text{max}} - 1$ images. This ensures that the inference step is computed on an input with $F_{\text{max}}$ images. During the meta-test phase, only the label which defines the prostate boundary is used. This is done to emulate the labels which may be available (via automatic segmentation) in practice with the application of the proposed method.

4.5 Experiments

4.5.1 Baseline Model Implementation and Training

The previously described meta-learning framework was implemented in TensorFlow [183] and Keras [184]. The implementation of the weakly-supervised image registration framework and corresponding loss functions used in this work were adapted from DeepReg (deepreg.net), an open-source Python package for medical image registration [247]. Hyper-parameters are all kept at defaults as described in [114] unless otherwise specified below. For data augmentation, each image-label pair was transformed by a random affine transformation, without flipping, before being input to the model during training.

The “Baseline” interactive registration model was trained for 250,000 iterations with the Adam optimizer [185], a minibatch size of 4, and an initial learning rate, $\beta^{\text{task}}$, of $1 \cdot 10^{-5}$. In the meta-training phase, the value of $k$ for task-level learning was 10, and the initial meta-learning rate, $\beta^{\text{meta}}$, was set to 0.5, with a linear decay to $1 \cdot 10^{-5}$ at the final training iteration. Loss weights $\gamma$ and $\alpha$ were both set to 1.0. $F_{\text{min}}$ is set as 2 and $F_{\text{max}}$ as 10. Requiring at least 2 frames allows the input to contain some spatial relationship between frames to help guide the registration. Training took approximately 120 hours on an NVIDIA DGX-1 system using a single Tesla V100 GPU. It is important to note that the number of iterations comprises each episode of task-level training, but does not include the meta-update; such is to say that when $k = 10$, a total of 25,000 (i.e. $\frac{250,000}{k}$) episodes of task-level learning (and subsequent meta-update) are performed, where each episode of task-level learning encompasses $k$ gradient updates.
4.5.2 Loss Functions

Two loss functions are employed to optimize the model parameters in training. In a weakly-supervised registration, the expected label similarity is maximized using a multiscale soft probabilistic DSC [114], which has been shown effectiveness especially when small foreground labels do not overlap initially. Substituting the interactive acquired TRUS labels $\ell_{mn}^{\text{target}}$ and the pre-operative MR labels $\ell_{mn}^{\text{source}}$:

$$L^*_{\text{sim}}(\phi) = \frac{1}{Z} \sum_{\sigma} S_{DSC} \left( f_{\sigma}(\ell_{mn}^{\text{target}}), f_{\sigma}(\ell_{mn}^{\text{source}}(\mathbf{u}_n^\phi)) \right),$$  \hspace{1cm} (4.11)

where $S_{DSC}$ is the soft probabilistic DSC [112], $f_{\sigma}$ is a 3D Gaussian filter with an isotropic SD $\sigma \in \{0, 1, 2, 4, 8, 16, 32\}$ in mm, and $Z = |\sigma|$. These published hyperparameter values [114] are used for comparison. In the proposed implementation, the deformation regulariser $L^*_{\text{def}}(\phi)$ that estimates bending energy [73] on $\mathbf{u}_n^\phi$ is used together with $L^*_{\text{sim}}(\phi)$ in Eq. (4.3) and Eq. (4.4).

4.5.3 Data

To train and evaluate the interactive registration model, a dataset comprising 108 pairs of pre-operative T2-weighted MR and intraoperative TRUS images from 76 patients which were acquired during the SmartTarget clinical trials [146] is used. Images were split into training and test sets comprising 88 and 20 patients, respectively, where no patient appears in both sets. Each of the MR and TRUS images were normalized to zero-mean and unit variance and resampled to an isotropic voxel size of $0.8 \times 0.8 \times 0.8$ mm$^3$. Prostate gland boundaries were segmented in the resampled MR and TRUS images. Segmentations of the prostate gland contours and landmarks in the MR images were acquired as part of the SmartTarget clinical trial protocols [146]. Additionally, segmentations of the prostate gland contours in the TRUS images were acquired automatically from the original TRUS slices [197], while the prostate gland landmarks were manually segmented. The data collection and annotation protocols for the data used in the below experiments are afore-described in Section 3.6.1.
4.5.4 Comparison with Meta-Learning Variants

Without extensively searching or refining all meta-learning hyper-parameters, which may lead to a misrepresentation of generalization capabilities, a sample of experimental results and validation on several variants of the proposed Baseline are provided. Each variant modifies one meta-learning hyper-parameter. Apart from the hyper-parameters which are specified to have changed in each variant described below, all hyper-parameters were kept fixed.

First, two of the proposed variants modify the number of gradient updates performed in each episode of task-level learning, \( k \), changing from the Baseline value of 10, to 1 and 100. It is of note that when \( k = 1 \), the Reptile algorithm corresponds to a single step of SGD on the expected loss [232]. As such, it may be interpreted that the \( k = 1 \) variant indicates jointly training the model on a mixture of all tasks, without meta-learning, before few-shot learning to obtain a fine-tuned model. Though \( k \) is often defined as \( \leq 10 \) in other meta-learning applications [232], training with a higher value of \( k \) is presented in the \( k = 100 \) variant, to demonstrate the resulting performance of a model which has been training with gradients that deviate greatly from those encountered in regular training. Due to the changes introduced to the training process (for \( k = 1 \)), and the deviation of the gradients from those which would normally be encountered in a non-meta-learning-based training protocol (for \( k = 100 \)), these variants are hypothesized to likely underperform relative to the baseline.

Second, a further two of the proposed variants use a modified initial meta-learning rate, \( \beta^{meta} \), changing from the Baseline value of 0.5 to 0.25 and 1.0. The linear decay remains unchanged, with a value for the meta-learning rate of \( 1 \cdot 10^{-5} \) at the final training iteration. An initial meta-learning rate \( \beta^{meta} \) which is too small was found to degrade performance as the resulting initial gradient steps may be uninformative, or overall, require additional training time, whereas a value that is too large may cause the gradients to deviate from those encountered in regular training. To prevent arbitrary selection, values that correspond closely with those initially presented with Reptile [232] were chosen.
Finally, the last two proposed variants modify the maximum number of frames used in training, $F_{\text{max}}$, to 5 and 15. The $F_{\text{max}}$ variants are primarily used to assess whether a different number of possible interactions or a different number of frames presented in training yields any difference in performance. It is expected that a higher and lower $F_{\text{max}}$ would result in better and worse performance, respectively. Though if the increase in performance gained per additional frame diminishes as the number of frames increases, training with a smaller $F_{\text{max}}$ may be beneficial. Conversely, if the increase in performance per additional frame does not significantly diminish, training with a higher $F_{\text{max}}$ and acquiring additional frames throughout the acquisition may be prudent in practice.

### 4.5.5 Comparison with State-of-the-Art Approaches

To demonstrate the effectiveness of the interactive meta-learning approach, it is compared to the application of a ‘registration’ without any initial alignment, and with a simple initialization whereby the prostate gland centroids are aligned. Furthermore, it is compared to two state-of-the-art approaches for deformable pairwise medical image registration; the weakly-supervised training approaches for LocalNet, described in [114], and VoxelMorph, described in [209].

In all comparisons, complete 3D volumes are utilized for source and target input images – unlike the proposed interactive meta-learning approach, which provides a sparse target input. Hyper-parameters are all kept at defaults as described in [114] and [209], and loss weights $\gamma$ and $\alpha$ are set to 1.0 in both instances.

### 4.5.6 Comparison with Non-Meta-Learning Approaches

To further demonstrate the effectiveness of the proposed interactive meta-learning approach, the sparse 2D target input of the proposed interactive meta-learning approach is emulated on the aforementioned State-of-the-Art approaches by training instances of LocalNet and VoxelMorph with 5 or 10 randomly sampled 2D target input images. Furthermore, the effects of the few-shot learning process used during adaptation of the meta-learning approach on these conventionally trained models and the meta-learning Baseline are also demonstrated by performing inference without
any few-shot learning. Lastly, to illustrate the effectiveness of the meta-learning approach to derive an initialization that may be rapidly adapted, LocalNet and VoxelMorph models are randomly initialized, and have few-shot learning applied to the untrained networks.

It is important to note that the impact of sparse data, as used throughout this series of comparisons, was not investigated by [114] or [209]. This means that it is unclear if the performance of these approaches will be adversely impacted by this reduction of data. However, this assessment may give perspective on the potential improvements and effects that a meta-learning- and few-shot learning approach may provide for image registration. It also provides a useful benchmark against which the performance of the proposed meta-learning approach may be gauged, where the comparison uses comparable amounts of input data.

4.5.7 Evaluation of Registration Methods

To compare the performance of the Baseline to all aforementioned methods, interactions are simulated which represent a clinically realistic scenario, on the real-world, clinical test data. This scenario reflects the continuous acquisition of frames through a right-to-left sweep through the prostate with the TRUS probe, obtaining a series of sagittal images that are uniformly distributed through the prostate (Figure 4.5). As noted in Section 4.4.7, two images are initially acquired, as $F_{\text{min}} = 2$, to make some spatial information between frames available in this first acquisition to help guide the registration.

The accuracy of the prostate surface point registrations was quantified using the DSC, and TRE; calculated as the distance between the 3D locations of corresponding, manually identified anatomical landmarks in the TRUS and MR image labels [114, 199]. All statistical tests which compare an evaluated method to the Baseline are based on two-tailed paired t-tests, at significance level $\alpha = 0.05$.

The DSC reported is computed between the warped MR label and the ground-truth label of the entire TRUS volume. The TRE is defined as the root-mean-square of each of the distances computed between the geometric centroids of the registered pairs of source and target landmarks for each patient. In the utilized
4.6. Results

Figure 4.5: Illustration of possible TRUS images acquired from the TRUS transducer in the presented clinical scenario. Newly acquired images (dashed lines) are captured in the sagittal plane (left) and are shown with the previously acquired images (solid lines) throughout one continuous ‘sweep’ through the prostate with the TRUS probe until full coverage of the prostate has been obtained. Figure from [207].

dataset, introduced in Section 3.6.5, the landmarks consisted of 309 pairs of points, including points defining the apex, base, urethra, visible lesions, junctions between the gland, gland zonal separations, vas deferens and the seminal vesicles, and other patient-specific point landmarks such as calcifications and fluid-filled cysts [146]. As mentioned in Section 3.6.5, such landmarks have been previously utilized to yield an overall spatial distribution that is representative of the full TRE distribution in this application [114, 118–120, 123–125, 130–133, 137, 139, 154, 157, 162, 167, 169, 170, 200–207], and permit this work to provide not only an evaluation of registration accuracy but to provide an estimate of the registration errors, such as those associated with tumor localization. For the Baseline model, the computational time per few-shot learning gradient update and subsequent registration in the meta-test phase are also reported.

4.6 Results

4.6.1 Baseline Performance

During few-shot learning in the meta-test phase, a gradient update and inference step for the Baseline model requires $0.67 \pm 0.07s$ and $0.37 \pm 0.05s$, respectively. Therefore, during adaptation, a fine-tuned task-specific Baseline model may be obtained, from which a predicted registration can be computed in approximately 6s
(8 gradient updates and 1 inference step). It is important to note that this computation takes place during acquisition, making it possible for such a method to be used in real-time in an iterative process whereby an image is acquired, segmented, and a gradient update is performed with this new data, until sufficient images are acquired. Given that the registration process occurs during acquisition, this is considerably less time than the 2 - 4 minutes required for image acquisition, contouring, and registration processes in conventional image-fusion targeted biopsies, such as those utilized in the SmartTarget clinical trials [146].

After 8 gradient updates of few-shot learning, the Baseline interactive registration model achieved a median TRE of 4.26 mm (mean = 5.06 mm) and a mean DSC of 0.85 were obtained with 10 input TRUS frames. With a median TRE of 4.26 mm, the registration performance obtained with the Baseline method is within range of previously defined clinically significant thresholds of 2.97 mm [208] and 5 mm [200]. Detailed results summarizing TRE and DSC throughout various steps of the few-shot learning process in the meta-test phase are given in Table 4.1. Example slices of input MR and TRUS image pairs and the registered MR images are provided in Figure 4.6 for qualitative visual assessment of the registration results for the Baseline at each step of few-shot learning in the meta-test phase, based on the test data.

<table>
<thead>
<tr>
<th>$F$</th>
<th>Gradient Updates</th>
<th>Mean TRE</th>
<th>Median TRE</th>
<th>Mean DSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>8.37 ± 4.08</td>
<td>7.02</td>
<td>0.77 ± 0.06</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>8.02 ± 3.98</td>
<td>6.98</td>
<td>0.79 ± 0.06</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>7.15 ± 4.17</td>
<td>6.02</td>
<td>0.81 ± 0.06</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>6.63 ± 4.11</td>
<td>5.61</td>
<td>0.82 ± 0.07</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>6.34 ± 4.16</td>
<td>5.34</td>
<td>0.82 ± 0.07</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>5.99 ± 4.08</td>
<td>5.27</td>
<td>0.83 ± 0.07</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>5.53 ± 4.12</td>
<td>4.34</td>
<td>0.84 ± 0.06</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>5.35 ± 4.13</td>
<td>4.37</td>
<td>0.84 ± 0.06</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>5.06 ± 4.19</td>
<td>4.26</td>
<td>0.85 ± 0.06</td>
</tr>
</tbody>
</table>
4.6. Results

Figure 4.6: Example image slice from one test case. The left-most column contains image slices from the source MR volume. The right-most column contains the corresponding target TRUS image slice. Other columns present the warped source MR image, resulting DDF, alternating vertical slices of the warped MR and target TRUS image, and warped MR prostate gland contour (Red) overlaid on the target TRUS prostate gland contour (Green), using the Baseline at a given shot of training during few-shot learning, with $F$ frames (e.g. the 2nd column indicates no few-shot learning, i.e. using the learned initialization with no fine-tuning, and two input frames, the 3rd column indicates one-shot of learning and three input frames, etc.). Figure from [207].

4.6.2 Performance of Baseline Variants

When varying $k$ in meta-training, after 8 gradient updates of few-shot learning, $k = 1$ had a median TRE of 4.48 mm (mean = 5.68 mm) and mean DSC of 0.83, whereas $k = 100$ had a median TRE of 4.58 mm (mean = 6.04 mm) and mean DSC of 0.85. A Shapiro-Wilk test was performed on the TRE and DSC values for each variant. In all instances, the test did not show evidence of non-normality ($p > 0.05$). In both variants, no significant difference was found between TRE ($p = 0.68$, $p = 0.46$) or DSC ($p = 0.58$, $p = 0.91$), based on two-tailed paired t-tests at $\alpha = 0.05$, relative to the Baseline. Detailed results summarizing the effects of different values of $k$ during training on TRE are illustrated in Figure 4.7 and summarized for TRE and DSC at each step of few-shot learning in the meta-test phase in Table 4.2.

When varying $\beta_{meta}$ in meta-training, after 8 gradient updates of few-shot learning, $\beta_{meta} = 0.25$ had a median TRE of 4.33 mm (mean = 5.80 mm) and mean DSC of 0.84, whereas $\beta_{meta} = 1.0$ had a median TRE of 3.29 mm (mean = 5.07 mm) and mean DSC of 0.87. A Shapiro-Wilk test was performed on the TRE and DSC values for each variant. In all instances, the test did not show evidence of non-normality ($p > 0.05$). In both variants, no significant difference was found
Figure 4.7: Tukey’s boxplots of TRE for the Baseline and all variants in the MR-TRUS registration experiment. Whiskers indicate 10th and 90th percentiles. Results are presented for registrations with 10 frames unless otherwise indicated. For $F_{\text{max}}$ variants, results are presented at 10 frames for direct comparison to the Baseline and other variants, and also with the number of frames corresponding to the value of $F_{\text{max}}$ used in training. Figure from [207].

between TRE ($p = 0.60, p = 0.85$) or DSC ($p = 0.74, p = 0.38$), based on two-tailed paired t-tests at $\alpha = 0.05$, relative to the Baseline. Detailed results summarizing the effects of varying $\beta^{\text{meta}}$ during training on TRE are illustrated in Figure 4.7 and summarized for TRE and DSC at each step of few-shot learning in the meta-test phase in Table 4.3.

When varying $F_{\text{max}}$ in meta-training, after $F_{\text{max}} - F_{\text{min}}$ gradient updates of few-shot learning, $F_{\text{max}} = 5$ had a median TRE of 4.50 mm (mean = 6.18 mm) and mean DSC of 0.85, whereas $F_{\text{max}} = 15$ had a median TRE of 3.58 mm (mean = 5.49 mm) and mean DSC of 0.84. A Shapiro-Wilk test was performed on the TRE and DSC values for each variant. In all instances, the test did not show evidence of non-normality ($p > 0.05$). In both variants, no significant difference was found between TRE ($p = 0.36, p = 0.82$) or DSC ($p = 1.00, p = 0.91$), based on two-tailed paired t-tests at $\alpha = 0.05$, relative to the Baseline.

Additionally, to yield a more direct comparison to the Baseline, after 8 gradient updates of few-shot learning, $F_{\text{max}} = 5$ had a median TRE of 4.44 mm (mean = 5.85
Table 4.2: Summary TRE and DSC for the $k = 1$ and $k = 100$ variants at each step of Few-Shot Learning in the Meta-Test phase in the MR-TRUS registration experiment. Mean values are presented ± SD. TRE is given in mm.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$F$</th>
<th>Gradient Updates</th>
<th>Mean TRE</th>
<th>Median TRE</th>
<th>Mean DSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>8.30 ± 3.75</td>
<td>7.79</td>
<td>0.76 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>8.15 ± 3.88</td>
<td>7.55</td>
<td>0.78 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>7.57 ± 3.97</td>
<td>7.03</td>
<td>0.79 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td>6.94 ± 3.90</td>
<td>5.79</td>
<td>0.80 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>4</td>
<td>6.80 ± 3.90</td>
<td>5.59</td>
<td>0.80 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>5</td>
<td>6.33 ± 4.12</td>
<td>5.49</td>
<td>0.80 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>6</td>
<td>5.83 ± 4.07</td>
<td>4.93</td>
<td>0.81 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>7</td>
<td>5.71 ± 4.01</td>
<td>4.43</td>
<td>0.83 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>8</td>
<td>5.68 ± 3.96</td>
<td>4.48</td>
<td>0.83 ± 0.05</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0</td>
<td>8.52 ± 3.86</td>
<td>7.83</td>
<td>0.76 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>8.42 ± 4.00</td>
<td>7.05</td>
<td>0.78 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>7.77 ± 4.12</td>
<td>6.49</td>
<td>0.79 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td>7.31 ± 4.20</td>
<td>5.88</td>
<td>0.81 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>4</td>
<td>7.04 ± 4.32</td>
<td>6.03</td>
<td>0.82 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>5</td>
<td>6.66 ± 4.43</td>
<td>5.64</td>
<td>0.83 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>6</td>
<td>6.27 ± 4.44</td>
<td>5.18</td>
<td>0.84 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>7</td>
<td>6.15 ± 4.48</td>
<td>4.78</td>
<td>0.84 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>8</td>
<td>6.04 ± 4.48</td>
<td>4.58</td>
<td>0.85 ± 0.04</td>
</tr>
</tbody>
</table>

mm) and mean DSC of 0.85, whereas $F_{\text{max}} = 15$ had a median TRE of 4.83 mm (mean = 6.30 mm) and mean DSC of 0.81. In both variants, no significant difference was found between TRE ($p = 0.56$, $p = 0.30$) or DSC ($p = 1.00$, $p = 0.31$), based on two-tailed paired t-tests at $\alpha = 0.05$, relative to the Baseline. Detailed results summarizing the effects of varying $F_{\text{max}}$ during training on TRE are illustrated in Figure 4.7, and summarized for TRE and DSC at each step of few-shot learning in the meta-test phase in Table 4.4.

Notably, the $F_{\text{max}} = 5$ variant performs better than the $F_{\text{max}} = 15$ variant for all values of $F \leq 5$. This is likely due to the distribution of the input images in the presented clinical scenario, whereby one continuous sweep of the prostate occurs, as presented in Figure 4.5. For example, when $F = 5$, while the input frames of the $F_{\text{max}} = 5$ variant will be evenly distributed across the entire prostate, while the 5 input frames of the $F_{\text{max}} = 15$ variant will be condensed into the right-most third of
Table 4.3: Summary TRE and DSC for the $\beta_{\text{meta}} = 0.25$ and $\beta_{\text{meta}} = 1.0$ variants at each step of Few-Shot Learning in the Meta-Test phase in the MR-TRUS registration experiment. Mean values are presented ± SD. TRE is given in mm.

<table>
<thead>
<tr>
<th>$\beta_{\text{meta}}$</th>
<th>$F$</th>
<th>Gradient Updates</th>
<th>Mean TRE</th>
<th>Median TRE</th>
<th>Mean DSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>2</td>
<td>0</td>
<td>8.08 ± 4.00</td>
<td>7.06</td>
<td>0.75 ± 0.07</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>7.90 ± 3.95</td>
<td>6.95</td>
<td>0.76 ± 0.07</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>7.33 ± 4.03</td>
<td>6.44</td>
<td>0.78 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td>6.80 ± 3.88</td>
<td>6.70</td>
<td>0.80 ± 0.06</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>4</td>
<td>6.49 ± 4.00</td>
<td>5.35</td>
<td>0.80 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>5</td>
<td>6.27 ± 4.04</td>
<td>5.26</td>
<td>0.81 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>6</td>
<td>5.97 ± 4.06</td>
<td>4.62</td>
<td>0.82 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>7</td>
<td>5.87 ± 4.11</td>
<td>4.31</td>
<td>0.83 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>8</td>
<td>5.80 ± 4.11</td>
<td>4.33</td>
<td>0.84 ± 0.05</td>
</tr>
<tr>
<td>1.0</td>
<td>2</td>
<td>0</td>
<td>8.37 ± 3.76</td>
<td>7.54</td>
<td>0.79 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>7.91 ± 3.77</td>
<td>7.17</td>
<td>0.81 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>6.88 ± 3.73</td>
<td>6.62</td>
<td>0.83 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td>6.21 ± 3.64</td>
<td>5.05</td>
<td>0.84 ± 0.05</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>4</td>
<td>5.88 ± 3.64</td>
<td>4.41</td>
<td>0.84 ± 0.04</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>5</td>
<td>5.53 ± 3.71</td>
<td>4.22</td>
<td>0.85 ± 0.04</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>6</td>
<td>5.27 ± 3.84</td>
<td>3.64</td>
<td>0.86 ± 0.04</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>7</td>
<td>5.13 ± 3.93</td>
<td>3.22</td>
<td>0.87 ± 0.04</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>8</td>
<td>5.07 ± 3.97</td>
<td>3.29</td>
<td>0.87 ± 0.04</td>
</tr>
</tbody>
</table>

the prostate, resulting in less spatial information being presented about the remaining prostate volume.

Example slices of input MR and TRUS image pairs and the registered MR images are provided in Figure 4.8 for qualitative visual assessment of the registration results, for each above-described variant, based on the test data.

4.6.3 Performance of State-of-the-Art Approaches

When applying no initial registration or alignment, a median TRE of 32.4 mm (mean = 36.1 mm) and mean DSC of 0.66 are obtained. Further, a median TRE of 18.4 mm (mean = 20.2 mm) and mean DSC of 0.77 are obtained if only prostate gland centroid alignment is performed on the images.

The performance of the Baseline model was not found to be significantly different, based on two-tailed paired t-tests at $\alpha = 0.05$, than LocalNet [114] for TRE and DSC ($p = 0.99, p = 0.36$), where a median TRE and mean DSC of 3.97
Table 4.4: Summary TRE and DSC for the $F_{max} = 5$ and $F_{max} = 15$ variants at each step of Few-Shot Learning in the Meta-Test phase in the MR-TRUS registration experiment. Mean values are presented ± SD. TRE is given in mm.

<table>
<thead>
<tr>
<th>$F_{max}$</th>
<th>$F$</th>
<th>Gradient Updates</th>
<th>Mean TRE</th>
<th>Median TRE</th>
<th>Mean DSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>0</td>
<td>7.90 ± 3.80</td>
<td>6.49</td>
<td>0.79 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>7.01 ± 3.94</td>
<td>5.67</td>
<td>0.82 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>6.34 ± 3.91</td>
<td>4.58</td>
<td>0.84 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td>6.18 ± 3.93</td>
<td>4.50</td>
<td>0.85 ± 0.04</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>0</td>
<td>8.10 ± 4.01</td>
<td>7.19</td>
<td>0.76 ± 0.04</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>7.98 ± 4.08</td>
<td>6.82</td>
<td>0.77 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>7.89 ± 4.19</td>
<td>6.53</td>
<td>0.78 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td>7.60 ± 4.27</td>
<td>6.33</td>
<td>0.79 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>4</td>
<td>7.08 ± 4.16</td>
<td>5.71</td>
<td>0.80 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>5</td>
<td>6.72 ± 4.14</td>
<td>5.51</td>
<td>0.81 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>6</td>
<td>6.54 ± 4.10</td>
<td>5.36</td>
<td>0.81 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>7</td>
<td>6.46 ± 4.06</td>
<td>5.44</td>
<td>0.81 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>8</td>
<td>6.30 ± 4.03</td>
<td>4.83</td>
<td>0.81 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>9</td>
<td>6.07 ± 3.99</td>
<td>4.37</td>
<td>0.82 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>10</td>
<td>5.78 ± 3.96</td>
<td>4.03</td>
<td>0.83 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>11</td>
<td>5.60 ± 3.95</td>
<td>3.86</td>
<td>0.84 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>12</td>
<td>5.52 ± 4.01</td>
<td>3.64</td>
<td>0.84 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>13</td>
<td>5.49 ± 4.00</td>
<td>3.58</td>
<td>0.84 ± 0.06</td>
</tr>
</tbody>
</table>

mm (mean = 5.27 mm) and 0.87 are obtained. Additionally, the performance of the Baseline model was not found to be significantly different, based on two-tailed paired t-tests at $\alpha = 0.05$, than VoxelMorph [209], for TRE and DSC ($p = 0.85$, $p = 0.47$), where a median TRE and mean DSC of 4.32 mm (mean = 5.62 mm) and 0.84 are obtained. A Shapiro-Wilk test was performed on the TRE and DSC values for each approach. In all instances, the test did not show evidence of non-normality ($p > 0.05$).

Detailed results summarizing the TRE of the Baseline and the non-meta-learning-based methods are illustrated in Figure 4.9. Example slices of input MR and TRUS image pairs and the registered MR images are provided in Figure 4.10 for qualitative visual assessment of the registration results, for each approach, based on the test data. It is important to note that these methods use complete 3D volumes for source and target input images, and achieves comparable performance to the
4.6. Results

**Figure 4.8:** Example image slices from one test case. The left-most column contains image slices from the source MR volume and the corresponding target TRUS image slice. Other columns present the warped source MR image, resulting DDF, alternating vertical slices of the warped MR and target TRUS image, and warped MR prostate gland contour (Red) overlaid on the target TRUS prostate gland contour (Green), using the above-labelled network; either the Baseline or one of its variants. Figure from [207].

<table>
<thead>
<tr>
<th></th>
<th>k = 1</th>
<th>k = 100</th>
<th>β_{meta} = 0.25</th>
<th>β_{meta} = 1.0</th>
<th>k_{max} = 5</th>
<th>k_{max} = 15</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td><img src="image" alt="Baseline k = 1" /></td>
<td><img src="image" alt="Baseline k = 100" /></td>
<td><img src="image" alt="Baseline β_{meta} = 0.25" /></td>
<td><img src="image" alt="Baseline β_{meta} = 1.0" /></td>
<td><img src="image" alt="Baseline k_{max} = 5" /></td>
<td><img src="image" alt="Baseline k_{max} = 15" /></td>
</tr>
<tr>
<td><strong>Target</strong></td>
<td><img src="image" alt="Target k = 1" /></td>
<td><img src="image" alt="Target k = 100" /></td>
<td><img src="image" alt="Target β_{meta} = 0.25" /></td>
<td><img src="image" alt="Target β_{meta} = 1.0" /></td>
<td><img src="image" alt="Target k_{max} = 5" /></td>
<td><img src="image" alt="Target k_{max} = 15" /></td>
</tr>
</tbody>
</table>

proposed method, which uses between two and ten frames of the target image in training and at inference. This represents between 1.6% and 8.5% of the complete 3D volume, which contains 118 image slices.

### 4.6.4 Performance of Non-Meta-Learning Approaches

When emulating the sparse target input of the proposed interactive meta-learning approach on LocalNet, though training without meta-learning, a median TRE of 7.51 mm (mean = 8.62 mm) and a mean DSC of 0.76 are obtained with 5 input images. Additionally, a median TRE of 6.26 mm (mean = 7.48 mm) and mean DSC of 0.79 are obtained with 10 input images. Performance of the fine-tuned Baseline model is significantly different than that observed when providing 5 and 10 inputs images,
4.6. Results

Figure 4.9: Tukey’s boxplots of TRE for the Baseline and all state-of-the-art methods in the MR-TRUS registration experiment. Whiskers indicate 10th and 90th percentiles. Figure from [207].

for TRE ($p < 0.01$, $p = 0.04$). However, no significant difference is observed with respect to DSC ($p = 0.07$, $p = 0.08$).

When using VoxelMorph, a median TRE of 7.36 mm (mean = 8.46 mm) and a mean DSC of 0.78 are obtained with 5 input images. Additionally, a median TRE of 5.86 mm (mean = 7.29 mm) and mean DSC of 0.81 are obtained with 10 input images. Performance of the fine-tuned Baseline model is significantly different than that observed when providing 5 inputs images, but not for 10 images, for TRE ($p < 0.01$, $p = 0.08$). However, no significant difference is observed with respect to DSC ($p = 0.09$, $p = 0.12$).

Applying few-shot learning to these same models at inference, LocalNet obtains a median TRE of 7.64 mm (mean = 8.83 mm) and mean DSC of 0.76 are obtained with 5 input images. Additionally, a median TRE of 7.23 mm (mean = 8.15 mm) and a mean DSC of 0.73 are obtained with 10 input images. VoxelMorph obtains a median TRE of 7.30 mm (mean = 8.74 mm) and a mean DSC of 0.79 are obtained with 5 input images. Additionally, a median TRE of 5.81 mm (mean = 7.33 mm) and a mean DSC of 0.81 are obtained with 10 input images. Overall, this illustrates that few-shot learning has minimal effects when applied to the conventionally trained
4.6. Results

Figure 4.10: Example image slices from two test cases, shown on the left and right, respectively. The left-most column contains image slices from the source MR volume and the corresponding target TRUS image slice. Other columns present the warped source MR image, resulting DDF, alternating vertical slices of the warped MR and target TRUS image, and warped MR prostate gland contour (Red) overlaid on the target TRUS prostate gland contour (Green), using the above-labelled network; either the Baseline or one of the state-of-the-art methods. Figure from [207].

LocalNet and VoxelMorph instances, without the meta-trained network initialization.

Using the Baseline, without any few-shot learning at inference, a median TRE of 4.57 mm (mean = 6.01 mm), and a mean DSC of 0.82 are obtained. While not significantly different, these values indicate poorer performance as compared to the Baseline when using few-shot learning. Together with the above results obtained when applying few-shot learning to the conventionally trained LocalNet and VoxelMorph, these results may indicate that the few-shot learning process is more effective on a meta-learned initialization.

Lastly, to assess the effects of the initialization, applying the few-shot learning process to an untrained model, where the weights are initialized randomly, results in
4.7. Discussion

This work presented a detailed description of the proposed deep-learning framework for meta-learning initializations for interactive medical image registration. Additionally, the registration, interaction, and meta-learning approach for the exem-
4.7. Discussion

Figure 4.12: Example image slices from one test case. The left-most column contains image slices from the source MR volume and the corresponding target TRUS image slice. Other columns present the warped source MR image, resulting DDF, alternating vertical slices of the warped MR and target TRUS image, and warped MR prostate gland contour (Red) overlaid on the target TRUS prostate gland contour (Green), using the above-labelled network; either the Baseline or one of the non-meta-learning-based methods. Figure from [207].

A clinical application: a multimodal volume-to-sparse prostate MR-TRUS image registration problem, are defined. Unlike conventional learning-based registration methods, where inadequate performance can occur given inconsistent image quality, varied imaging protocols, or interpatient variation; the use of interactions assists the proposed method in predicting a more accurate solution. By introducing additional data at inference, the proposed method is permitted to refine its predictions in real-time, thereby compensating for deficiencies that may be infeasible to address during training, while learning a patient-specific registration model. This is made all the
4.7. Discussion

more effective by framing the proposed method as a gradient-based meta-learning problem. This formalizes the pre-training-then-fine-tuning pathway for improving fine-tuning capabilities with fast adaptation to new data at inference.

As illustrated in Fig. 9, the performance of the Baseline network for volume-to-sparse of the proposed method provides registration accuracy which is not statistically different from the compared 3D-to-3D method and is significantly lower than the compared volume-to-sparse 3D-to-2D methods. Furthermore, analysis of Baseline variants indicates that the proposed method is not highly sensitive to changes in the defined meta-learning hyper-parameters, as demonstrated by the absence of statistical difference between the reported accuracies of all variants. The highly comparable and/or improved performance of the proposed method, compared to non-meta-learning methods, demonstrates the flexibility and generalizability obtained from a wide range of hyper-parameters illustrate the usefulness of interactive registration methods and meta-learning. These improvements provide the basis to enable the use of such a framework in other medical image registration applications.

A point of critical importance for multimodal image registration, when using intensity-based methods, is the lack of clear voxel-level spatial correspondence between pairs of images. To overcome this challenge, the weakly-supervised registration method described by [114], used as the underlying registration method in this work, employs corresponding labeled structures during training, while requiring no labels at inference. However, as few-shot learning is utilized during the meta-test phase, real-time prostate gland segmentations are required. High DSC values and rapid inference times have been reported for single slice prostate gland segmentations from TRUS images [118, 197], and previously adopted into end-to-end registration frameworks [118]. As such, the need for a method that generates TRUS prostate gland segmentations must be considered when implementing such a registration framework in clinical practice, but should not be considered prohibitive to the real-time implementation of interactive registration in practice given that the addition of these additional segmentation inference steps would add, at most, several seconds to the total time required to compute the registration.
4.8. Conclusion

Furthermore, it is worth noting that both the proposed interactive registration and the 3D-to-2D methods, presented in Section 4.5.6, require the positional information for all TRUS images relative to each other, or a fixed reference to establish spatial correspondence between images. This positional information permits the creation of a volume comprising each frame, to which the MR volume may be registered. In practice, this spatial information may be obtained using some form of positional tracking; where the probe is affixed to a stepper for mechanical tracking, or where the probe is used freehand with electromagnetic or optical tracking. Without this positional information, the current implementation is unlikely to successfully predict a suitable registration. Further experiments are required to assess the suitability of the proposed framework for registration of MR to un-tracked TRUS images, however, this is considered out of the scope of this work, where it was sought to compare the 3D-to-sparse-2D method directly to the widely-researched 3D-to-3D registration.

4.8 Conclusion

This chapter has presented a novel meta-learning-based framework for interactive medical image registration. In summary, the work has demonstrated one such exemplar application through partial registration of MR to sparsely acquired intra-operative TRUS images. This method obtains similar registration accuracies to state-of-the-art 3D image registration methods which require complete image volumes. Additionally, this method significantly outperforms those same state-of-the-art methods when applied to the same challenging partial data problem. This demonstrates the effectiveness and efficiency of the proposed real-time interactive MR to partial US image registration method, which may be applied during intraoperative procedures, such as prostate biopsy.
Chapter 5

Test-Time Meta-Registration Optimization

This chapter is based on the work entitled “Meta-Registration: Learning Test-Time Optimization for Single-Pair Image Registration”, published in MICCAI ASMUS Workshop 2022 [248].

5.1 Introduction

“Classical” pairwise approaches pose the image registration problem as an optimization for transformation, which maximizes a given image similarity measure between the transformation-warped source image and the target image. Much work has been dedicated to variants in transformation models, similarity metrics and optimization algorithms [225]. While these classical methods are usually applied to a single pair of images, recent learning-based methods utilize deep neural networks to predict the transformation, or simply a DDF, between any source and target images. Typically, these networks are optimized with a set of pairs of training images, minimizing a loss function that is based on image similarity measures or distance between corresponding segmentations [249–251]. Other works [207, 215] have proposed to use meta-learning to adapt registration networks to new types of images, with a distinct aim of efficient intra- or inter-domain adaptation.

More recently, deep neural networks have also been proposed to represent, or parameterize, the spatial transformation between a single pair of images. This
becomes analogous to classical methods; permitting the network to be optimized “without training data” [252]. In this chapter, this single-pair optimization process is considered as an iterative optimization as opposed to a learning-based problem where the phrase “learning without training data” may be used. Thus, it follows that this same optimization of a single pair of images may then also be applied to improve a registration network obtained from the learning-based methods as a case of test-time optimization [217], and not only from the original source and target images. Both single-pair optimization approaches have been shown to improve existing methods which use learning-based registration networks alone. This may be due to the use of networks or data which are prone to overfitting, perhaps due to sensitivity to initialization, limited available training data or highly variable clinical imaging, and sometimes to underfitting due to over-constrained transformations.

Observed from these prior studies, both the single-pair methods, including those using neural networks, and the learning-based methods may have advantages in seeking pair-specific features and population-statistics-based features that are useful to align the image pair of interest. As such, this chapter proposes the use of meta-learning to combine population-based, generalizable learning and single-pair optimization, by considering image pairs in training as different meta-tasks. This allows the meta-training to optimize a meta-registration network that can be effectively and efficiently adapted to individual test image pairs, using single-pair test-time optimization.

This is particularly useful for registering US images. US often creates challenging registration tasks with clinically acquired data given their known high variability and varying quality, due to user- and view-dependency. The registration of US to US images to other images can help to create a more complete and accurate picture of the area being imaged. This is particularly useful in the context of prostate imaging, where the prostate gland is often imaged from different angles and at different times. Through registration, clinicians can more easily compare and analyze the images to identify any changes or abnormalities. In longitudinal or change detection scenarios, where changes in the size or shape of the prostate gland, and possible
lesions, may need to be measured accurately over time, ensuring that measurements are consistent and accurate, allowing for more reliable tracking of any changes in the prostate gland. US to US registration can also be especially useful in the context of population-based assessments, as it allows for the aggregation and analysis of data from multiple patients to identify trends or patterns. Additionally, the generation of interpatient atlases can provide a reference for comparing and analyzing individual patient images, allowing for more accurate and consistent diagnoses and treatment planning in the future. In all instances, a rapid optimization process can be critical in enabling real-time image guidance in potentially many surgical and interventional applications, as well as for image analysis more broadly. In this work, 3D US images obtained from TRUS-guided prostate cancer interventions are used to demonstrate the feasibility, accuracy, and speed of single image-pair optimization using the proposed meta-registration algorithm.

5.2 Contributions

This chapter describes a meta-learning and network training paradigm for combining population-based, generalizable learning and single-pair optimization, by considering image pairs in training as different meta-tasks. In doing so, this work demonstrates the effectiveness of meta-registration methods as compared to strictly generalizable, conventional learning-based methods and single-pair optimized classical iterative methods.

The below sections provide a thorough description of unsupervised learning-based image registration, a meta-registration framework, and the test-time single-pair optimization process which are applied to the exemplar application of unimodal image registration on US images. Subsequent validation and analysis of registration performance are presented and compared to different learning-based methods and classical iterative methods.
5.3 Methods

5.3.1 Unsupervised Learning-Based Image Registration

This section describes the proposed meta-registration using an unsupervised loss, as outlined in Figure 5.1. Similarly to a general purpose learning-based image registration method, as defined in Section 4.3.1, when given \( N \) pairs of training source and target images, \( \{x_{source}^n\} \) and \( \{x_{target}^n\} \), where \( n = 1, \ldots, N \), existing approaches predict the voxel correspondence \( u_n^\phi = f^\phi(x_{source}^n, x_{target}^n) \), i.e. the transformation that aligns the two images, using a registration network \( f^\phi \) with network parameters \( \phi \). Where this differs, is that for an unsupervised learning algorithm, the training goal thus is minimizing a loss function over \( N \) training pairs, to obtain the optimal \( \phi^* \):

\[
\hat{\phi} = \arg\min_\phi \sum_{n=1}^N \left[ L_{sim}(\phi | x_{source}^n, x_{target}^n) + \alpha^\phi L_{def}(\phi | x_{source}^n, x_{target}^n) \right],
\] (5.1)

where \( L_{sim}(\phi | x_{source}^n, x_{target}^n) = L_{sim}(x_{source}^n(\phi^\phi), x_{target}^n) \) is a negative image-based similarity measure, a function between the transformation-warped source images \( x_{source}^n(\phi^\phi) \) and the target images \( x_{target}^n \), and \( L_{def}(\phi | x_{source}^n, x_{target}^n) = L_{def}(u_n^\phi) \) is the deformation regularization, encouraging the smoothness of the transformation \( u_n^\phi \) weighted by a hyperparameter \( \alpha^\phi \). Unlike the proposed interactive registration framework from Section 4, a negative weak-supervision loss based on label similarity is not added, given the unsupervised nature of the method, where only image-level similarity is used to guide learning.

As such, during test time with an unseen pair of images, \( x_{source}^{test} \) and \( x_{target}^{test} \), the trained network \( f^\phi \) predicts the transformation that aligns the two, \( u^\phi_{test} = f^\phi(x_{source}^{test}, x_{target}^{test}) \).

5.3.2 Test-Time Single-Pair Optimization

Consider an optimization problem to align a pair of test images \( x_{source}^{test} \) and \( x_{target}^{test} \):
Figure 5.1: Schematic representation of the proposed unsupervised meta-registration method for single-pair test-time optimization. A learning-based registration is trained over multiple episodes during the training phase (left). In each episode, a pair of images is sampled and repeatedly registered. Following each episode, the meta-update updates the registration model based on the learned gradients from the episode which was just completed. Once training is complete, the registration model may be optimized at test-time for a single pair of images (right) using few-shot learning to yield a registration model optimized for a specific pair of input images. Figure from [248].

\[
\hat{\theta} = \arg\min_\theta \left[ \mathcal{L}_{\text{sim}}(\theta \mid x_{n,\text{source}}, x_{n,\text{target}}) + \alpha_{\theta} \mathcal{L}_{\text{def}}(\theta \mid x_{n,\text{source}}, x_{n,\text{target}}) \right], \quad (5.2)
\]

where \( \alpha_{\theta} \) is the deformation hyperparameter. This is equivalent to the classical pairwise registration, iteratively optimizing a transformation network \( f_{\theta} \) with its randomly initialized parameters \( \theta \), which (re-)parameterize the transformation \( u_{n} = f_{\theta}(x_{n,\text{source}} \mid x_{n,\text{test}}, x_{n,\text{target}}) \) between \( x_{n,\text{test}}^{\text{source}} \) and \( x_{n,\text{test}}^{\text{target}} \).

Alternatively, when the parameters are initialized by the trained registration network parameters, \( \theta = \hat{\phi} \), as obtained in Eq. (5.1) and Eq. (5.2), represents test-time optimization for the given test pair.

It is also noteworthy that the transformation network \( f_{\theta} \) could be a different network to the registration network \( f_{\phi} \), whilst this study uses a single network to facilitate a model-agnostic implementation of the proposed meta-registration algorithm.
5.3. Methods

5.3.3 Model-Agnostic Meta-Learned Test-Time Optimization

This section describes the proposed meta-registration algorithm. Each pair of images is considered a different meta-task, such that a meta-training scheme can be adopted to improve the test-time optimization. During the meta-training, different meta-tasks are sampled. The resulting bi-level optimization thus becomes:

$$\phi^* = \arg\min_{\phi} \sum_{n=1}^{N} \left[ L_{\text{sim}}(\theta \mid x_n^{\text{source}}, x_n^{\text{target}}, \theta^{*}(n)(\phi)) + \alpha^{\phi} L_{\text{def}}(\phi \mid x_n^{\text{source}}, x_n^{\text{target}}, \theta^{*}(n)(\phi)) \right], \quad (5.3)$$

s.t. $$\theta^{*}(n)(\phi) = \arg\min_{\theta} \left[ L_{\text{sim}}(\theta \mid x_n^{\text{source}}, x_n^{\text{target}}, \phi) + \alpha^{\theta} L_{\text{def}}(\theta \mid x_n^{\text{source}}, x_n^{\text{target}}, \phi) \right], \quad (5.4)$$

where, the outer optimization in Eq. (5.3) obtains the optimum meta-parameters $$\phi^*$$, such that $$\theta^{*}(n)(\phi)$$ is an optimized network for individual $$n^{th}$$ tasks. In the proposed meta-registration, $$\theta$$ and $$\phi$$ are shared network parameters. Therefore, model-agnostic meta-learning algorithms such as MAML [231] or Reptile [232] can be readily applied to solve this bi-level optimization problem.

The proposed meta-registration may be considered by two different views of combining the learning-based method and the test-time optimization:

1. it optimizes a learning-based registration network that can be used for better test-time optimization;

2. it is an iterative method for registering a single pair of images, using a neural network to parameterize the spatial transformation, which can be initialized with prior knowledge learned from training data.

It is also interesting to note that data augmentation methods may be considered as the samples of individual tasks in the proposed meta-registration.
5.4 Experiments

5.4.1 Meta-Registration Implementation

Reptile [232] is adopted as the gradient-based meta-learning strategy employed in the meta-registration framework as it provides a computationally efficient optimization of the gradient-update procedure. Reptile was designed to quickly learn to perform a new task with minimal training, which suits the test-time single-pair optimization process. This is achieved in practice through a bi-level optimization. In the inner optimization loop, an episode of task-level learning is applied over \( k \) mini-batches. In the outer optimization loop, SGD is performed by using the difference between the model weights prior to and after the inner optimization loop’s episode of task-level learning.

The meta-learning methodology described in this work adapts a learning-based registration method available from the unsupervised image registration framework within DeepReg [247]. This ‘Baseline’ meta-registration model architecture utilizes LocalNet [114], and was trained for 200,000 iterations with the Adam optimizer [185], a mini-batch size of 4, and an initial learning rate of \( 1 \cdot 10^{-5} \). Through the meta-training phase, the value of \( k \) used was 10, with an initial meta-learning rate, \( \beta^{meta} \), of 0.5, linearly decaying to \( 1 \cdot 10^{-5} \) over the course of the 200,000 iterations. The SSD loss as \( L_{sim} \) and bending energy [73] as \( L_{def} \). The deformation hyperparameter \( \alpha_{def} \) was set to 10.0 to weight the deformation regularization relative to the image similarity loss. During the inner optimization, data augmentation is applied to the source and target images. Each image was independently transformed by a random affine transformation, without flipping, prior to being used as input. Training required approximately 120 hours on an NVIDIA DGX-1 system using a single Tesla V100 GPU.

In the meta-test phase, test-time optimization is performed via few-shot learning with 5 gradient updates on the sampled pair of test images. This yields a test-time optimized registration model which can perform accurate registrations on the test data. In this optimization process, a mini-batch size of 1 is used, and 5 gradient updates are performed to fine-tune the model. Apart from these values, the few-shot
5.4. Experiments

learning uses the same hyperparameters as in the inner optimization loop during the meta-training phase.

5.4.2 Data

To train and evaluate the meta-registration, 108 intraoperative TRUS images from 76 patients, acquired during the SmartTarget clinical trials [146], are utilized. The TRUS images were split into a training set and a test set, with each comprising 88 and 20 images, respectively, where no patient appears in both sets. Images were normalized and resampled to an isotropic voxel size of $0.8 \times 0.8 \times 0.8 \text{mm}^3$. The TRUS segmentations of the prostate gland boundary were acquired automatically [197], and any additional landmarks used to compute registration accuracy, such as apex and base, were segmented manually. The data collection and annotation protocols for the data used in the below experiments are afore-described in Section 3.6.1.

5.4.3 Comparison Studies

To demonstrate the effectiveness of the meta-registration approach it is compared to a classical iterative non-rigid registration method, and two state-of-the-art architectures for deformable medical image registration [114, 209]. Additionally, the effects of the test-time optimization are demonstrated by comparing it to the meta-registration baseline without any few-shot learning.

The meta-registration method is first compared to a conventional iterative registration approach, whereby SGD is applied over 3,000 iterations to directly learn a DDF which describes the transform between a given pair of source and target images. Here, a learning rate of 0.01 is applied, and use the same loss and deformation hyperparameter as in the training of the meta-registration method. Subsequently, the meta-registration method is compared to two widely-used approaches, LocalNet [114] and VoxelMorph [209], for deformable pairwise medical image registration using unsupervised learning. In both instances, these networks are trained with identical loss, training, and deformation hyperparameters to the meta-registration method. To illustrate the effects of the test-time optimization process, as well as demonstrate the effectiveness of the meta-learned initialization, a comparison to
only the meta-learned initialization without any test-time optimization or fine-tuning applied is presented.

5.4.4 Evaluation of Registration Methods

The accuracy of the prostate surface registrations was quantified using the DSC, and TRE; calculated as the distance between the 3D locations of corresponding, manually identified anatomical landmarks in the TRUS images [114, 199]. Reported DSC are computed between the transformed prostate gland label of the source image and the ground-truth prostate gland label of the target image. TRE is reported as the root-mean-square of the distances between landmark centroids of the pairs between the transformed source image and the target image. The computational time required at inference, on GPU, is reported for each method.

5.5 Results and Discussion

During few-shot learning in the test-time single-pair optimization process, a gradient update and inference step require approximately 0.67s and 0.37s, respectively. Therefore, during the test-time optimization, the meta-registration method requires approximately 3.7s to be fine-tuned and provide a prediction for the specific image pair. This is notably much less than the classical method evaluated by nearly 100 times, while delivering comparable performance. Conversely, this 3.7s is nearly 10 times slower than other existing and evaluated learning-based methods which do not use any test-time optimization.

While requiring an additional 3s compared to other existing learning-based methods, the performance of DSC and TRE is significantly improved. This performance is significantly different with respect to DSC and TRE from LocalNet and VoxelMorph, based on two-tailed paired t-tests, at a significance level of $\alpha = 0.05$.

After 5 gradient updates of few-shot learning through the meta-test phase, the Meta-Registration method yields a mean TRE of 6.1 mm and a DSC of 0.74. This is comparable, but improved, to the Meta-Registration initialization which is learned during meta-training, which yields a mean TRE of 6.2 mm and a DSC of 0.73. Both of these results are also comparable to the Classical Non-Rigid optimized
5.6 Conclusion

This chapter has presented a meta-registration framework for test-time single-pair optimization of ultrasound images. Results that were comparable to time-consuming, classical iterative methods were obtained in a fraction of the time. Additionally, the meta-registration method outperforms existing learning-based methods with minimal additional time required during inference for the test-time optimization process. These results demonstrate a critical step in enabling adaptive, tailored real-time image guidance in many surgical and interventional applications.
Figure 5.2: Tukey’s boxplots of DSC for all methods. Whiskers indicate 10th and 90th percentiles. Figure from [248].

Figure 5.3: Tukey’s boxplots of TRE for all methods. Whiskers indicate 10th and 90th percentiles. Figure from [248].
Figure 5.4: Example image slices from one test case. The left-most column contains image slices from the source and target images. Other columns present the warped image, a checkerboard of the warped and target images, the warped prostate gland contour (Red) overlaid on the target prostate gland contour (Green), and the resulting DDF using the above-labelled method. Figure from [248].
Chapter 6

Conclusion

In this thesis, several novel approaches which provide fast, accurate, and generalizable medical image registration have been presented. The development of such methods illustrates multiple avenues by which the data efficiency and user interactivity may be increased and optimized for different registration tasks within a broad sample of target application domains.

Chapter 3 introduced FPT, a learning-based approach for point-set registration. This approach demonstrated the ability to learn effective individual point displacements without the need for point correspondence in several synthetic, non-medical domains as well as in US-based spine atlas reconstructions and in MR-TRUS image registration. In all instances, FPT was shown to be robust to deformation, noise, and, more saliently, robust to partial input data. This ability to effectively and efficiently manage partial data problems opens the door to solving other medical image registration problems where the absence of complete training data is a limiting factor for existing solutions.

Chapter 4 introduced a meta-learning-based framework for interactive medical image registration. In MR-TRUS image registration using sparsely acquired intra-operative TRUS images, similar registration accuracies to state-of-the-art full-volume methods were obtained. Notably, when compared to the same state-of-the-art methods trained only with partial data, the interactive method significantly outperforms. This provides further evidence of the data efficiency and effectiveness of an interactive registration approach for scenarios where real-time registration is an
important enabling technology, as in prostate biopsy.

Chapter 5 introduced a second meta-learning-based framework for test-time, single-pair optimization of US images. Compared to the classical, iterative methods for single-pair optimized image registration, the proposed solution was found to achieve comparable or more accurate registration with a large reduction in computational time required owing to the efficient and adaptive test-time optimization process. Overall, the proposed method demonstrates a critical advance in learning-based registration methods, which are often entirely iterative or entirely reliant on a single prediction, towards time-efficient registration tailored to specific clinical applications.

The validation data presented for the deep-learning-based medical image registration methods presented in this thesis indicate fast, accurate, data-efficient, and patient-specific registrations. Given the common challenges with the often limited size of datasets available in medical imaging, the data-efficient approaches presented are of interest for further study in other related applications where registration is required. Furthermore, given other inherent challenges with medical image registration, especially in multimodal applications where non-linear relationships exist between image intensities, the ability to learn descriptive, data-driven features directly from limited amounts of partial data without compromising registration accuracy presents several new directions for research through the application of application-specific constraints or interactions. Additionally, while the ability of the presented methods to adapt to new patient populations or varied image intensities remains untested, the potential for such methods to do so presents powerful opportunities for future work. In summary, this thesis represents significant progress towards generally-applicable, data-efficient, and adaptable learning-based non-rigid registration methods.

6.1 Future Work

While image registration itself is not a recently developed concept, the application of deep-learning-based methodologies to image registration is relatively new. There are several key advantages to deep learning approaches – generally, and in medical
image registration specifically – namely the rapid inference times, comparable or improved accuracy to classical or conventional methods, and the ability to derive novel interpretations and features from the data. Such advantages have made real-time image registration a clinical possibility, especially when paired with real-time imaging modalities, such as US, unlike some classical, iterative methods. Despite the success of many deep learning methods, the inherent reliance on data availability and the potential lack of robustness on new, unseen data – as compared to classical methods – is problematic. Without the underlying data (or the ability to appropriately simulate sufficiently realistic data) for a specific application or task, it may not be possible to even attempt a given problem when using a deep learning approach.

Several aspects of the work presented in this thesis have demonstrated the promise of methods such as weak supervision, user interaction, and the use of partial data for medical image registration. However, there are still several challenges which need to be addressed before these methods can be used in a clinical setting. The following sections outline some of these challenges; such as the difficulties in determining appropriate regularization/constraints, defining effective user interactions, and working with datasets of limited size as avenues for future work.

### 6.1.1 Learning to Constrain ‘Free’ Transformations

In Chapter 3, a ‘model-free’ approach for point-set registration is introduced. The rationale for introducing such a method is due in part to the difficulty associated with adequately selecting an effective parametric transformation model that does not explicitly define constraints on spatial coherence or smoothness. Often such constraints need to be made explicit and are hand-engineered to be capable of handling noise, outliers, and missing data. This makes it difficult to apply methods that use these constraints to real-world data. In current practice, the key to devising effective transformation models, or to not using an inherent model, is to strike a pragmatic balance between exploiting expert/domain knowledge versus generating ‘knowledge’ using data. This can be thought of more generally as what ‘we’ know about the required transformation and data versus what a model may learn from large quantities of high-quality labeled data.
Given the sparseness of the zonal structures within the prostate, it has been previously argued that smoothness constraints are required [114]. In intensity-based registrations, not imposing physically plausible constraints on registration processes can result in highly distorted local deformations, leading to poorer registration accuracy, or a higher TRE. Heuristically designed deformation regularisations, such as those discussed in Section 1.2.3.3 [73, 89, 90], have demonstrated their benefits in various intensity-based registration applications, including the work done in this thesis. However, in feature-based registration methods, the regularization process is not always as well defined. This is due to the potential dependence on generating plausible deformation from either smoothing or a finite element model- [131, 136, 162, 180, 253] or statistical model-based [118, 135, 162] approaches. While these methods have been applied in practice, spatial smoothing may be considered too simple to model prostate motion, and the generation of finite element or statistical models can be cumbersome and computationally intensive.

Recently, the physics-informed neural network (PINN), a deep learning-based method capable of embedding a knowledge of physics described by partial differential equations, has gained popularity by virtue of its ability to direct and constrain its learning process through prior knowledge of a series of specified physical laws [254, 255]. In doing so, PINNs have the potential to maximize the information content available within the training data by leveraging the embedded prior physical knowledge in order to better generalize with less data [254, 255].

With PINNs, it may be possible to utilize existing learning-based methods and additionally embed biomechanical constraints through a system of equations that are solved alongside the registration loss in an end-to-end manner. By training in this way, a generalizable, learned method for complex or elastic constraints may be developed, as has been shown in cardiac [256–258] and neuroimaging [259] applications. This would remove the need to constrain registrations with the aforementioned finite element- or statistical-based models or to develop registration algorithms with no constraints as in the work of Chapter 3, which may lead to physically implausible deformations. The use of PINNs may improve the registration accuracy and quality by
permitting learned deformations that are constrained differently in different regions of the prostate that exhibit different biomechanical properties. They also ensure that any large deformations are consistent and realistic in the context of prostate motion. In practice, the FPT framework, presented in Chapter 3, may be extended to regress not only the per-point displacements but additional information or inputs for various required constraint or regularization loss functions. Such information or loss functions may include the prediction of stresses or the computation of a loss on the elastic energy based on the predefined properties at a given point.

Furthermore, because these rigidity constraints may be set – or learned – explicitly on a per-point basis, giving every point in the point-set a variable amount of rigidity, the use of PINNs within the FPT framework could be extended to the quantification of spine curvature in the registration of US-based reconstructions to the generic spine models, presented in Section 3.5. While previous works have been predominantly intensity-based, implementing different forms of rigidity or biomechanical constraints for surgical guidance using a combination of hand-engineered heuristics [260, 261], recent works have adopted learning-based methods. These learning-based methods have been used to regress a series of local level-wise transformations between each vertebra [262] or to leverage weak supervision to constrain the transformation by penalizing the overlap and DDF with localized rigid transformation within the vertebral bodies [263]. Such methods, therefore, require the decisions made with respect to the actual resulting deformation to be explicitly set, rather than learned within a set of physics-informed guiding parameters and formulae to determine the optimal registration.

Summarily, adopting learned biomechanical constraints into the FPT framework, or otherwise, may provide a means to improve the registration accuracy and quality by permitting learned deformations that are constrained differently in different regions of the prostate that exhibit different biomechanical properties. This can benefit patients through the reduction of TRE and MR-TRUS registration errors, which may lead to more accurate biopsy targeting and treatment planning. Furthermore, the use of PINNs within the FPT framework could be extended to the quantification of spine
curvature in the registration of US-based reconstructions to the generic spine models, presented in Section 3.5. This may provide a means to improve the accuracy of the registration of US to the generic spine model, which may lead to more accurate quantification of spinal curvatures.

6.1.2 Expanding Definitions of Interactivity

Throughout Chapter 4, an interaction was considered to be any action taken by the user or model which has a reciprocated action taken by another entity – either the user or model – as a result of the initial action. This permitted a formulation for interactions in such a way that meant they do not have to be used independently. From the two categories defined in Section 4.3.2, computer-to-user and user-to-computer, and looking beyond the broad scenarios defined in Section 4.4.1, we may further gain an understanding of the practical and potential uses for each type of interaction.

Overall, the core differences between these two broad categories may best be described by assessing which entity is ‘waiting’ to take an action. In user-to-computer interactions – the category most similar to that considered in Chapter 4 – an example of a more classically interactive scenario is given. Here the model waits for new data before being able to make its next prediction and then makes a new prediction as the user continues their task of adding data or error correction. Conversely, in computer-to-user interactions, the model is often making its best initial prediction, where it may continue to accept new information from the user should the user think it is required on the basis that it may improve or refine the initial prediction.

Another perspective from which to consider these differences is to compare the ‘type’ of feedback given in each scenario. In user-to-computer interactions, typically the interactions will illustrate a type of positive feedback where the user continually provides a ‘positive’ form of interaction to the model. This may be in the form of new data or additional annotations which can assist in the registration process. On the other hand, in computer-to-user interactions, the interactions which may occur may be considered more akin to ‘negative’ feedback. This negative interaction may be in
6.1. Future Work

the form of identifying mistakes that the model has made or replacing poor-quality images so that the model may be able to derive an improved prediction from ‘better’ data. Given these positive and negative considerations and criteria, as well as for the purposes of defining additional interactions, it may be helpful to consider the user-to-computer or computer-to-user interactions in a given application as error correction or as the introduction of new data. In the following, each of these options and their possible implementations and accompanying implications are discussed.

With error correction, an interaction may be viewed as an instance where the model takes an action (i.e. makes a prediction), and the user responds by providing a reciprocated action. It is expected that this action would be in the form of an error correction, or, perhaps, the confirmation of a successful registration as demonstrated by the inability to identify errors. It follows that error correction interactions may permit image re-acquisition or annotation within areas that are not well aligned, as in Chapter 4, where image acquisition occurs on a local (i.e. one or a few images) level. It is also possible that this acquisition may occur on a global level (i.e. over the entire image volume) when some of the initially acquired images are of insufficient quality or there has been significant patient motion. While image re-acquisition may be most effectively used in tandem with real-time imaging, such as US, given that image quality may be operator dependent and that images may be rapidly re-acquired, this does not preclude volumetric acquisitions being omitted from consideration for re-acquisition interactions. Volumetric image re-acquisition may be especially effective as part of a rapid adaptation and registration process in scenarios where interventional MR imaging is in use, as is needed for guidance in ablation- or radiation therapy-based interventions for liver [264] and prostate cancers [265–267] or for neurosurgical guidance [268, 269], where the registration of pre-operative and intra-operative imaging is crucial.

Additional error-correcting interactions may include user-defined annotations which can be used to indicate regions in the registration that are not well-aligned. Such a scenario is more likely to apply when image registration is performed outside of the operating room, and with reduced or negligible time pressure, potentially
in a similar manner to existing interactive segmentation tools, such as MONAI Label [270], where the user continues to delineate 3D images to actively re-train a deployed segmentation model through their annotations.

While the processes mentioned above are inherently different methods from which an interactive registration model may be learned, both of these error-correction interactions may be formulated similarly. For example, the interactions may be used to direct learning by only modifying the annotated or re-scanned regions of the registration by applying constraints to the possible deformations instead of modifying the entire registration. Such modifications may be implemented as constraints on the transformation or deformation field, or perhaps as different weights for given voxels or images for any image- or label-based loss functions, or otherwise.

Looking at ‘addition-of-new-data’ interactions, we may consider these actions to follow a process whereby the user takes an action (i.e. providing new data to the model), and the model provides a reciprocated action by giving a refined prediction. Therefore, the addition of new data may include interactions such as image acquisition, annotation of any well-aligned regions, or landmark segmentation. As mentioned with error-correction interactions, these different types of interactions may occur on a local or global level when additional data or annotation is thought necessary by the user to predict an acceptable registration.

User-defined annotations, in contrast to those defined for error correction, may also be used to indicate to the model which areas of the registration are visibly well aligned. Additionally, user-defined or automatic segmentations of landmarks or anatomical regions – and the manual confirmation of those regions – may be used to indicate areas or structures which should be aligned to the model. This may be especially useful in multimodal registration tasks where there are different intensities between the pair of images being registered. The acquisition of new images may be formulated as an interaction that modifies the models’ input data, whereas the previously discussed interactions may be implemented similarly to error-correction interactions, either as constraints on the resulting transformation or deformation field or by differently weighting any image- or label-based loss function at the voxel-
or image-level. In doing so, the interactions may be able to direct the model to preserve the transformation or deformation field in regions that are well aligned and only modify the registration elsewhere. Further, the addition of segmentations may also be used as part of a label-similarity-based loss function to promote overlap between corresponding labels in the images being registered in a supervised test-time optimization approach.

6.1.3 Enabling Open and Reproducible Science in Multimodal Medical Image Registration

The fusion of clinically important information, such as pre- and intra-operative imaging, as is achieved by registering MR and TRUS images, remains important to many surgical and interventional tasks. While the registration of such images can assist in the delivery of prostate biopsy and focal therapy, as discussed and as motivated in this thesis, and has transformed the standard of prostate cancer patient care, there remain ongoing challenges with access to large quantities of high-quality data (and any associated annotations) and the production of open and reproducible science.

The data used in this thesis were not publicly accessible. Often, the data which are openly available to the research community are often scarce and of low quality, and frequently lack expert annotations. The process of annotation can be time-consuming and costly, requiring extensive expert domain knowledge and experience. While crowd-sourced labels have been used in previous works [271–273], the quality of such labels is often unknown and may be inconsistent. Further, the use of crowd-sourced labels may be limited to tasks where the labels are easily defined, such as in image segmentation, and may not be easily applied to other tasks, such as image registration.

As such, an ongoing major contribution of this thesis is the provision of well-curated, expert-annotated, real-world data for research use – available through [274] through the organization and delivery of the ‘MR to Ultrasound Registration for Prostate Challenge’ (µ-RegPro) [275] at the 26th International Conference on Medical Image Computing and Computer Assisted Intervention. The µ-RegPro
6.1. Future Work

Challenge yields the release of one of the first multimodal medical imaging datasets, complete with expert annotations for validation, for benchmarking advancement in registration methodologies, as well as for future research in managing the most common non-skin cancer in men.

The µ-RegPro Challenge will provide access to the MR and TRUS data of over 100 subjects with mp-MRI who underwent an MR-targeted TRUS-guided prostate biopsy procedure to assist in the diagnosis and staging of their prostate cancer as part of the SmartTarget Biopsy Clinical Trial [146].

This data will permit the µ-RegPro Challenge to evaluate the performance of multimodal image registration methods between pre- and intra-operative imaging methods for surgical and interventional tasks. Using intensity-based methods, feature-based methods, or some combination of the two, participants will provide a function that accepts as input a:

- target image,
- source image,
- source label.

The function must also produce as output a:

- warped source image,
- DDF.

Participants are permitted to deform or transform the images in any manner (e.g. parametric or non-parametric transformation) which they should choose, however; they must provide an equivalent DDF for purposes of metric computation on the test data. To assist with benchmarking and to provide a baseline for state-of-the-art performance, participants are provided with end-to-end weakly supervised solutions based on LocalNet [114] and VoxelMorph [209].

As previously described, the overarching goal for the µ-RegPro Challenge is to provide high-quality, open research data for the community. Though organized as a single event in conjunction with 26th International Conference on Medical Image
Computing and Computer Assisted Intervention that does not repeat yearly, it is anticipated that the data will provide long-term value to the community with the prospect of additional challenges and data being organized and released in the future.
Bibliography


[100] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In


[123] Tharindu De Silva, Aaron Fenster, Derek W. Cool, Lori Gardi, Cesare Romagnoli, Jagath Samarabandu, and Aaron D. Ward. 2d-3d rigid registration


Pingkun Yan, Sheng Xu, Ardeshir R Rastinehad, and Brad J Wood. Adversarial image registration with application for mr and trus image fusion. In


[193] David Malfair, Anne K Flemming, Marcel F Dvorak, Peter L Munk, Alexandra T Vertinsky, Manraj K Heran, and Doug A Graeb. Radiographic evaluation


[200] Vaishali V Karnik, Aaron Fenster, Jeffrey Bax, Derek W Cool, Lori Gardi, Igor Gyacskov, Cesare Romagnoli, and Aaron D Ward. Assessment of im-


[206] Sharmin Sultana, Daniel Y Song, and Junghoon Lee. A deformable multimodal image registration using pet/ct and trus for intraoperative focal prostate


[221] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin,


[228] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional


[243] Peter A Pinto, Paul H Chung, Ardeshir R Rastinehad, Angelo A Baccala, Jochen Kruecker, Compton J Benjamin, Sheng Xu, Pingkun Yan, Samuel


[263] Bailiang Jian, Mohammad Farid Azampour, Francesca De Benetti, Johannes Oberreuter, Christina Bukas, Alexandra S Gersing, Sarah C Foreman, Anna-


[274] Zachary M. C. Baum, Shaheer U. Saeed, Zhe Min, Yipeng Hu, and Dean C. Barratt. Mr to ultrasound registration for prostate challenge - dataset, 2023.

[275] Zachary M.C. Baum, Shaheer U. Saeed, Zhe Min, Yipeng Hu, and Dean C. Barratt. Mr to ultrasound registration for prostate challenge, 2023.